

(19) 日本国特許庁 (J P)

## (12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2003-122669

(P2003-122669A)

(43) 公開日 平成15年4月25日 (2003.4.25)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テーマコード* (参考)
G 0 6 F 13/00	5 4 0	G 0 6 F 13/00	5 4 0 R 5 B 0 7 5
17/30	1 1 0	17/30	1 1 0 F
	2 2 0		2 2 0 Z
	4 1 9		4 1 9 B
17/60	1 4 8	17/60	1 4 8
審査請求 未請求 請求項の数10 OL (全 27 頁)			

(21) 出願番号 特願2001-314993(P2001-314993)

(22) 出願日 平成13年10月12日 (2001.10.12)

(71) 出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番  
1号

(72) 発明者 津田 宏

神奈川県川崎市中原区上小田中4丁目1番  
1号 富士通株式会社内

(74) 代理人 100074099

弁理士 大曾 義之 (外1名)

Fターム(参考) 5B075 ND03 ND08 ND14 ND36 NR01  
PQ40 PQ46

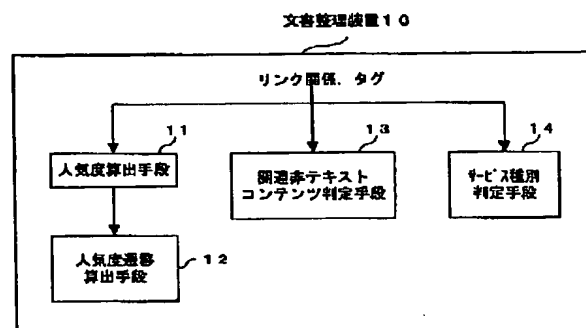
(54) 【発明の名称】 リンク関係に基づく文書整理方法

(57) 【要約】

【課題】 文書の人気度が増加する一方で減少することがないという問題を解決しつつ、文書が時系列的にどのような状況にあるのかを示す情報を得る事を可能とする。

【解決手段】 文書整理装置10は、人気度算出手段11及び人気度遷移算出手段12を備える。人気度算出手段11は、第1の期間に収集されたネットワーク上の文書間のリンク関係に基づいて、各文書の人気の高さの度合いを示す人気度を算出する。人気度遷移算出手段12は、第2の期間内に人気度算出手段11によって算出された人気度に基づいて、人気度の変化の方向と度合いを示す人気変化度を算出する。

## 本 発 明 の 原 理 図



【特許請求の範囲】

【請求項1】 ネットワーク上の文書の人気の高さの度合いである人気度を算出する制御をコンピュータに実行させるプログラムであって、  
文書からリンク関係を抽出し、  
第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、  
前記抽出された各文書の人気度を算出する、  
ことを含む処理を前記コンピュータに実行させることを特徴とするプログラム。

【請求項2】 前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出する、  
ことを更に含む処理を更にコンピュータに実行させることを特徴とする請求項1に記載のプログラム。

【請求項3】 第2の期間内に算出された前記人気度に基づいて、前記人気変化度を算出する、  
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項2に記載のプログラム。

【請求項4】 前記第2の期間内に算出された前記人気度の時間に対する回帰式を算出し、  
前記人気変化度を前記回帰式に基づいて算出する、  
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項3に記載のプログラム。

【請求項5】 前記回帰式の回帰係数に基づいて前記人気変化度を決定する、  
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項4に記載のプログラム。

【請求項6】 前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定する、  
ことを更に含む処理を前記コンピュータに実行させる事を特徴とする請求項4に記載のプログラム。

【請求項7】 ネットワーク上の文書間の関係を判定する制御をコンピュータに実行させるプログラムであって、

第1の文書からリンク関係を抽出し、  
前記リンク関係に基づいて、前記第1の文書からリンクされる第2の文書が、前記第1の文書の内容に関連する関連非テキスト文書であるか否かを判定する、  
ことを含む処理を前記コンピュータに実行させる事を特徴とするプログラム。

【請求項8】 ネットワーク上の文書が提供するサービスの種別を判定する制御をコンピュータに実行させるプログラムであって、  
前記文書からユーザ入力を指定するタグを抽出し、  
前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定する、  
ことを含む処理を前記コンピュータに実行させる事を特徴とするプログラム。

【請求項9】 ネットワーク上から文書を検索する文書検索方法であって、

前記ネットワークから文書を収集し、  
前記文書からリンク関係を抽出し、  
第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、  
前記抽出された各文書の人気度を算出し、  
検索条件に基づいて文書を検索し、  
前記検索された文書を前記人気度に基づいてランキングし、  
前記ランキング結果に基づいて、前記検索された文書に関する情報を出力する、  
ことを含むことを特徴とする文書検索方法。

【請求項10】 第2の期間内に算出された前記人気度に基づいて、前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出し、  
前記人気変化度に関する情報を前記検索された文書に関連する情報に加える、  
ことを更に含むことを特徴とする請求項9に記載の文書検索方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、ネットワーク上に存在する文書の整理に関し、特に文字情報のみならず、画像、音声等の様々な形態の大量の文書が存在し、かつそれらの文書が激しく変化するような場合に好適な文書整理技術に関する。

【0002】

【従来の技術】例えば、WWW (WorldWide Web、以下、ウェブという)は、急成長しているインターネットリソースである。ウェブには、2000年において20億ページ以上という調査があるように、大量の文書(ウェブページともいう)が存在する。また、ウェブには、存在する文書が大量であるだけでなく、非常に文書の変化が速いという特徴もある。

【0003】ウェブ・アーカイブ・オーガニゼーション(Web Archive Organization)による調査では、ウェブでは、情報が毎月10%ずつ増加し、一つの文書の寿命(文書が作成されてからメンテナンスされなくなるまで)は約75日という結果もある。

【0004】現在、このようなウェブ上に存在する情報を検索する検索サービスが、いくつか提供されている。この検索サービスにおいて、検索の結果得られた文書のネットワーク上の位置を示す情報、例えば、URI (Uniform Resource Identifier) 又はURL (Uniform Resource Locator)、とそのウェブページの内容を説明する文が、検索者に提供される。

【0005】また、近年、ブロードバンド時代を反映し、文書のコンテンツはテキストから動画・音声等に、また単に内容閲覧させる文書からサービスを提供する文書に、文書の内容が移行している。

【0006】

【発明が解決しようとする課題】しかし、従来の検索サービスでは、ある時点でのウェブの状況に基づいて検索サービスを提供しているため、文書が時系列的にどのような状況にあるか、例えば人気が出始めであるのか、定番的なものであるか、人気が落ちているものであるのかは不明であるという問題があった。例えば、ウェブから「最近人気のあるウェブページ」を調べる方法はなかった。

【0007】また、ウェブの場合、古くなった文書を作者が削除したり、文書の内容をこまめに更新したりすることはあまりない。そのため、単純に文書へリンクしている他の文書の数（被リンク数）に基づいて、文書の人気の高さの度合い、つまり人気度を算出すると、人気度が減るということは殆んどないという問題もあった。

【0008】また、ブロードバンド時代を反映して文書が、テキスト中心から、画像などの非テキストやサービスを含んだものが中心になっているが、その変化に対応した文書の整理方法がなかった。

【0009】以上の問題を鑑み、単純な被リンク数に基づく文書の人気度が増加する一方で減少することがないという問題を解決することを1つの目的とする。また、文書の人気度が時系列的にどのような状況にあるのかを示す情報を得る事を可能とすることを更なる目的とする。また、文書の内容等の移行に対応して文書を整理する事を可能とすることを更なる目的とする。

【0010】

【課題を解決するための手段】本発明の1態様によれば、ネットワーク上の文書の人気の高さの度合いである人気度を算出する人気度算出方法において、文書からリンク関係を抽出し、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出することを特徴とする。

【0011】第1の期間内に収集された文書又は第1の期間内に更新された文書を対象として人気度を算出することにより、古い文書を人気度を算出する対象から省き、ひいては、文書の人気度が増加する一方で減少することがないという問題を解決する。第1の期間は、有意な人気度を算出するために、ある程度長い期間、例えば150日程度であることが望ましい。

【0012】ここで、前記リンク関係及び前記文書の前記ネットワーク上の位置を示す文書位置情報に基づいて前記人気度を算出することとしてもよい。これにより、文書の内容を見ることが不要であるため、人気度を迅速に算出することが可能となる。

【0013】上記方法において、更に、第2の期間内に算出された前記人気度に基づいて、前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出することとしてもよい。これにより、文書の人気度が時系列的にどのような状況にあるのかを示す情報を得る事が可能となる。

【0014】ここで、第2の期間は、人気度の変化を見るために、あまり長い期間でない、例えば数週間程度である事が望ましい。上記方法において、前記第2の期間内に算出された前記人気度の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出することとしてもよい。この場合、前記回帰式の回帰係数に基づいて前記人気変化度を決定することとしてもよいし、前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定することとしてもよい。

【0015】また、回帰式を算出する際に、人気度の代わりに、前記抽出された文書の人気度に基づく順位を用いることとしてもよい。また、本発明の別の1態様によれば、ネットワーク上の文書間の関係を判定する文書関係判定方法において、第1の文書からリンク関係を抽出し、前記リンク関係に基づいて、前記第1の文書からリンクされる第2の文書が、前記第1の文書の内容に関連する非テキスト文書であるか否か判定することを特徴とする。これにより、近年多くなっている、画像など非テキストメディアの種別に応じて、文書を整理することが可能となる。

【0016】上記方法において、前記第1の文書から前記第2の文書にリンクする部分の近辺にある文字列を前記第1の文書から抽出し、前記文字列に基づいて、前記第2の文書が前記第1の文書の内容に関連する関連非テキスト文書であるか否か判定することを更に含むこととしても良い。例えば、文字列が、MEPG、動画、ストリーミング等、第2の文書が非テキストフォーマットであることを示す文字列である場合、第2の文書は第1の文書の内容に関連する非テキスト文書であると推定できる。

【0017】また、上記方法において、前記拡張子が特定の拡張子でない場合、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書でないことと決定することを含むこととしてもよい。拡張子は、第2の文書の文書フォーマットを示すため、これに基づいて非テキスト文書であるか否か判定する事ができる。

【0018】また、上記方法において、前記第2の文書が前記第1の文書内で所定回数以上使用されているか否かに基づいて、前記第2の文書は前記第1の文書の内容に関連する非テキスト文書であるか否か判定することとしてもよい。例えば、プリント等は、画像であるが、これらの文書作成用の素材系の画像は、1つの文書中で何度も繰り返して使用されることが多いため、第1の文書中での使用回数が多い第2の文書は第1の文書の内容に関連していないと推定する事が可能である。

【0019】また、上記方法において、前記第1の文書内に前記第2の文書のファイル名と類似したファイル名を持つ第3の文書がある場合、前記第2の文書の前記ファイル名が前記第3の文書の前記ファイル名よりも辞書順に若くない場合、前記第2の文書を第1の文書の内容

に関連する非テキスト文書としてデータベースに登録しないことを更に含むこととしてもよい。

【0020】例えば、第1の文書が写真集である場合、多くの画像を含む。これらの画像を全て第1の文書の内容に関連する非テキスト文書として登録すると、かえって煩雑となる可能性がある。しかし、この場合、画像ファイルのファイル名が互いに類似している事が多いため、複数の文書のファイル名のうち最も辞書順にファイル名が若い文書のみを第1の文書の内容に関連する非テキスト文書として登録することにより、このような煩雑さを解消する事ができる。

【0021】また、上記方法において、前記第2の文書からリンクされる第3の文書がある場合、前記第1の文書の前記ネットワーク上の位置を示す文書位置情報と前記第2の文書の文書位置情報に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定することを更に含むこととしても良い。また、前記第1の文書の前記文書位置情報と前記第3の文書の文書位置情報に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定することを更に含むこととしてもよい。

【0022】例えば、第1の文書中には、バナー広告等、文書の内容に関係ない非テキスト文書が第2の文書として含まれることがある。このような場合、前記第2の文書の前記文書位置情報と第2の文書のリンク先である第3の文書の前記文書位置情報が、前記第1の文書の前記文書位置情報と同じサーバアドレス又はドメインを持たないことが多いため、各文書の文書位置情報に基づいて、広告バナーのような第1の文書の内容に関連しない非テキスト文書を除く事ができる。

【0023】また、本発明の更なる別の1態様によれば、ネットワーク上の文書が提供するサービスの種別を判定するサービス種別判定方法において前記文書から、ユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定することを含むことを特徴とする。これによっても、近年の文書の内容等の変化に対応して、より具体的には、文書が提供するサービスの種別に応じて文書を整理することが可能となる。ユーザ入力を指定するタグとして、例えば、文書を記述する言語がHTMLである場合、フォームタグが挙げられる。

【0024】上記方法において、前記文書にユーザ入力を指定するタグが含まれていない場合、前記文書はサービスを提供しないと決定することを更に含むこととしてもよい。文書中に何もユーザ入力欄が含まれていない場合、その文書がサービスを提供している可能性は低いからである。

【0025】また、文書に含まれるボタンの表示に基づいて、前記文書が提供するサービスの種別を判定することを更に含むこととしてもよい。また、さらに、ボタン

の表示に加えて入力欄に基づいて、前記文書が提供するサービスの種別を判定することを更に含むこととしてもよい。文書が提供するサービスによって、多くの場合、ボタン等の入力欄の形式が決まっているからである。

【0026】より具体的には、例えば、前記文書に、商品を購入する旨を示す表示をもつボタンが含まれている場合、前記文書が提供するサービスの種別を販売店として決定することを更に含むこととしてもよい。商品を販売するサービスを提供する文書において、商品の注文を受けるために、このようなボタンが含まれる事が多いからである。

【0027】また、例えば、前記文書に、ユーザ入力エリア、及び検索を示す表示をもつボタンが含まれている場合、前記文書が提供するサービスの種別を検索として決定することを更に含むこととしてもよい。

【0028】また、本発明の各態様にかかわる方法において行われる手順を実現する手段を備える装置によっても、前述した方法と同様の作用・効果を得ることが可能である。また、上述した本発明の各方法において行なわれる手順と同様の制御をコンピュータに行なわせるプログラムをコンピュータに実行させる事によっても、前述した方法と同様の作用・効果を得ることが可能である。また、上述のプログラムを記録したコンピュータ読み取り可能な記録媒体から、そのプログラムをコンピュータに読み出させて実行させることによっても、前述した方法と同様の作用・効果を得ることが可能である。

【0029】

【発明の実施の形態】以下、本発明の実施の形態を図面に基づいて説明する。図1に、本発明の原理を示す。本発明に係わる文書整理装置は、リンク関係に基づいて、文書の人気の高さの度合いを示す人気度を算出し、さらに、その人気度が時系列的にどのように変化しているのかを示す人気変化度を算出する。そして、各文書を算出された人気度及び人気変化度に基づいて整理する。

【0030】図1に示すように、文書整理装置10は、人気度算出手段11、人気度遷移算出手段12を備える。人気度算出手段11は、第1の期間に収集されたネットワーク上の文書間のリンク関係に基づいて、各文書の人気の高さの度合いを示す人気度を算出する。ここで、人気度算出手段11は、第1の期間内に収集された文書又は第1の期間内に更新された文書を対象として人気度を算出する。これにより、文書の人気度が増加する一方で減少することがないという問題を解決する。

【0031】人気度遷移算出手段12は、第2の期間内に人気度算出手段11によって算出された人気度に基づいて、人気度の変化の方向と度合いを示す人気変化度を算出する。なお、人気度遷移算出部12は、人気変化度を算出する際に、人気度の代わりに、人気度に基づいて各文書をランキングした人気度順位を用いることとしても良い。これにより、ネットワーク上の文書の人気度が

時系列的にどのように変化しているのか解析する事が可能となる。

【0032】また、近年、ブロードバンドインターネット時代を反映して、文書の内容（コンテンツ）はテキストから画像、動画、音声等のような非テキストに、さらに、単に情報を読ませる文書から検索や登録などサービスを提供する文書に重点が移りつつある。しかし、例えば、従来の検索サービスにおいて、検索結果として、文書のネットワーク上の位置を示す情報とその文書の内容を説明する文とを検索者に提供するだけでは、その文書がどのような非テキストコンテンツを含んでいるのか、あるいは、その文書でどのようなサービスを行っているかは、その文書にアクセスしない限り、その検索者にはわからない。

【0033】また、このような非テキストコンテンツを整理する際に、単純にファイルの拡張子に基づいて文書に含まれる非テキストコンテンツを判定すると、その文書に含まれているバナーやブリット（点）など、その文書の内容とは関連のない非テキストコンテンツもその文書と関連するコンテンツとして整理されてしまうという問題もある。

【0034】そこで、図1に示すように、本発明に係わる文書整理装置10は、更に、関連非テキストコンテンツ判定手段13及びサービス種別判定手段14とを備える。関連非テキストコンテンツ判定手段13は、文書間のリンク関係に基づいて、各文書に含まれる非テキストコンテンツのうち、その文書の内容に関連する非テキストコンテンツを判定し、文書の内容に関連すると判定された非テキストコンテンツをその文書に対応させて整理する。

【0035】サービス種別判定手段14は、各文書に含まれるタグ、例えば入力欄を作成する際に用いるユーザー入力指定するタグ、例えば、HTMLの場合のフォームタグ等に基づいて、その文書がサービスを提供しているか否かを判定し、更に文書がサービスを提供している場合そのサービスの種別を判定し、判定したサービス種別をその文書に関連させて整理する。これにより、例えば、検索サービスにおいて、検索結果として、文書のネットワーク上の位置を示す情報とその文書の内容を説明する文に加えて、その文書の内容と関連する非テキストコンテンツ及びその文書で提供されているサービスについての情報を、その文書に関する情報として提供する事が可能となる。

【0036】以下、本発明の実施形態について説明する。なお、上述の文書整理装置をネットワーク上から文書を検索する文書検索装置に適用した場合について説明するが、本発明の適用範囲を限定する趣旨ではない。

【0037】図2に、本発明の実施形態に係わる文書検索装置の構成を示す。文書検索装置100は、ネットワークから文書を収集し、収集された文書を整理する。ネ

ットワークとして、イントラネットや専用回線等のLAN (Local Area Network)、公衆回線やインターネット等のWAN (Wide Area Network) が考えられる。文書検索装置100は、直接又は、不図示のネットワークを介して接続された端末（不図示）のユーザからの指示に従って、文書を検索し、検索結果をユーザに提供する。

【0038】なお、文書検索装置100がネットワークを介して端末にサービスやデータを提供するサーバである場合、ユーザの端末はブラウザ108を備え、ユーザは、ブラウザ108を用いて文書検索装置100から送信される情報を閲覧することとしてもよい。

【0039】図2に示すように、文書検索装置100は、収集部101、人気度算出部102、人気度遷移算出部103、関連非テキストコンテンツ判定部104、サービス種別判定部105、ページ分類部106、検索サービス部107、文書テーブル11、リンク関係テーブル112、人気度テーブル113、人気度変化テーブル114、非テキストコンテンツテーブル115及びサービス種別テーブル116を備える。収集部101、人気度算出部102、人気度遷移算出部103、関連非テキストコンテンツ判定部104、サービス種別判定部105、ページ分類部106及び検索サービス部107は、例えば、プログラムにより記述されたソフトウェアコンポーネントに対応し、文書検索装置100を実現するコンピュータのメモリの特定のプログラムコードセグメントに格納される。

【0040】ここで、ネットワーク上に存在する文書、つまりウェブページを記述する言語として、例えば、HTML (HyperText Markup Language)、XHTML (eXtensible HyperText Markup Language)、XML (eXtensible HyperText Markup Language)、SGML (Standard Generalized Markup Language) 等のような、リンク関係を文書内に埋め込む事が可能な言語が考えられる。また、本発明では、上記のような言語で記述されたテキスト文書以外に、画像、動画、音声等も文書として扱う。以下、テキスト文書を記述する言語をHTMLであると仮定して説明する場合もあるが、本発明を限定する趣旨ではない。

【0041】収集部101は、ネットワーク上で公開されている文書を収集し、収集された文書に、文書を識別する文書ID (IDentification information) を付す。さらに、収集部101は、収集された文書のリンク関係を解析する。そして、収集部101は、収集された文書のネットワーク上の位置を示す文書位置情報を文書テーブル111に格納し、収集された文書間のリンク関係に関する情報をリンク関係テーブル112に格納する。

【0042】ここで、文書位置情報として、例えば、URI (Uniform Resource Identifier) 等が考えられる。なお、URIは包括的な概念であり、現在は、URIの機能の一部を具体的に使用したURL (UniformRes

source Locator) が広く利用されている。以下、文書位置情報をURLであると仮定して説明する場合もあるが、本発明を限定する趣旨ではない。

【0043】人気度算出部102は、定期的(又は不定期的に)に、収集部101によって収集された文書のリンク関係に基づいて、人気の高さの度合いを示す人気度を算出し、算出結果を人気度テーブル113に格納する。人気度を算出する際に、人気度算出部102は、収集部101によって収集された文書のうち、第1の期間内に収集された文書又は第1の期間内に更新された文書を、人気度を算出する対象となる文書とする。ここで、第1の期間は、あまり短期間では人気度として意味のある結果を得る事ができないため、ある程度長い期間である必要がある。例えば、第1の期間として、人気度を算出する日の前150日間が考えられる。

【0044】これにより、作成された後、更新されずに放置されたままの文書を人気度を算出する対象から省く事が可能となる。延いては、ある文書の人気度を単純に時系列に算出すると、人気度が単調に増加する一方であるという問題を解決する事ができる。

【0045】人気度遷移算出部103は、第2の期間内に人気度算出部102が算出した人気度に基づいて、各文書について人気度の変化の方向と度合いを示す人気変化度を算出し、算出結果を人気度変化テーブル114に格納する。ここで、第2の期間は、あまり長いと、短期的な人気度の変動を把握する事ができないため、ある程度短い期間、例えば、数週間程度である必要がある。例えば、第2の期間として、人気変化度を算出する日の前14日間が考えられる。

【0046】より具体的には、例えば、人気度遷移算出部103は、各文書について、第2の期間内に算出された人気度を、人気度テーブル113から取得し、取得された人気度の時間に対する線形回帰式を算出し、その線形回帰式の回帰係数を人気変化度として得る。また、人気度遷移算出部103は、人気変化度を算出する際に、人気度の代わりに、人気度に基づいて各文書をランキングした人気度順位を用いることとしても良い。これにより、ネットワーク上の文書の人気度が時系列的にどのように変化しているのか解析する事が可能となる。

【0047】関連非テキストコンテンツ判定部104は、各文書の文書位置情報に含まれるファイル名の拡張子や、文書中のリンクが埋め込まれた部分の前後にある文字列に基づいて、各文書のタイプを判定する。更に、関連非テキストコンテンツ判定部104は、文書間のリンク関係に基づいて、各文書に含まれる非テキストコンテンツが、各文書の内容に関連するかどうか判定する。そして、関連非テキストコンテンツ判定部104は、各文書の内容に関連すると判定された非テキストコンテンツを、その文書に対応させて非テキストコンテンツテーブル115に格納する。これにより、各文書に含まれる非

テキストコンテンツのうち、その文書の内容に関連しない非テキストコンテンツを除去し、その文書の内容に関連する非テキストコンテンツを文書に対応させて整理することが可能となる。

【0048】サービス種別判定部105は、各テキスト文書に含まれる入力欄を記述する情報に基づいてその文書で提供するサービスの種別を判定し、判定されたサービス種別をその文書に対応させてサービス種別テーブル116に格納する。これにより、各文書が提供するサービスの種別を文書に対応させて整理することが可能となる。

【0049】ページ分類部106は、関連分野等に基づいて各文書を分類する。文書の分類方法については、既に様々な分類技術が存在するため、本実施形態では詳しく説明することを省略する。

【0050】検索サービス部107は、ユーザの指示に従って、ネットワーク上の文書を検索し、検索結果をユーザに提供する。その際に、検索サービス部107は、検索の結果得られた文書に関する情報を、人気度テーブル113及び人気度変化テーブル114から取得し、検索された文書の内容を説明する情報及び文書位置情報に加えて、人気度、人気変化度をユーザに提供する。これにより、ユーザは、検索された文書の人気が、今どのような状態にあるのか、人気が出始めてあるのか、人気落ちてきているのか、検索結果の出力画面で提供される情報によって知ることができる。

【0051】さらに、検索サービス部107は、検索の結果得られた文書に関する情報を、非テキストコンテンツテーブル115及びサービス種別テーブル116から取得し、検索された文書の内容に関連する非テキストコンテンツに関する情報及び検索された文書で提供されているサービス種別に関する情報もユーザに提供することとしてもよい。これにより、ユーザは、検索の結果得られた文書が、どのような非テキストコンテンツを含むのか、或いは、その文書でどのようなサービスが提供されているのか、その文書にアクセス(閲覧する)しなくとも、検索結果の出力画面で提供される情報によって知ることが可能となる。

【0052】また、ユーザが、1以上の文書の人気度に関する情報を提供するように要求した場合、検索サービス部107は、その文書に関する情報を人気度テーブル113、人気度変化テーブル114等から取得し、取得された情報を時系列に提供することとしてもよい。これにより、ユーザは、ある文書の人気度の推移を分析することが可能となる。

【0053】以下、図3から図8を用いて、各テーブルのデータ構造について説明する。まず、図3を用いて文書テーブル111のデータ構造について説明する。図3に示すように、文書テーブル111は、各文書について文書位置情報とそれに対応する文書IDを格納する。こ

れにより、各文書の文書位置情報は文書IDに変換され、以降の処理では文書IDを用いて各文書のリンク関係等に関する情報を管理することが可能となる。

【0054】次に、図4を用いて、リンク関係テーブル112のデータ構造について説明する。リンク関係テーブル112は、各文書についてのリンク関係情報を格納する。図4に示すように、リンク関係情報は、その文書が収集された日時（又は日付）、更新された日時（又は日付）、リンク元となっている文書の文書ID、リンク先となっている文書の文書IDを項目として含む。以下の説明において、リンク元となっている文書の文書IDをリンク元IDといい、リンク先となっている文書の文書IDをリンク先IDということとする。なお、各文書の更新日時が取得困難な場合、収集日時を更新日時に代えて扱う事としてもよい。

【0055】次に、図5を用いて、人気度テーブル113のデータ構造について説明する。人気度テーブル113は、各文書についての人気度情報を格納する。図5に示すように、人気度情報は、人気度が算出された日時（又は日付）、その文書の文書ID、算出された人気度及び、人気度に基づいて文書をソートした結果である人気度順位を項目として含む。

【0056】次に、図6を用いて、人気度変化テーブル114のデータ構造について説明する。人気度変化テーブル114は、各文書について人気度変化情報を格納する。人気度変化情報は、その文書の文書ID、人気度について線形回帰式を算出した結果得られた回帰係数（傾き）及び切片、並びに、人気度順位について線形回帰式を算出した結果得られた回帰係数（傾き）及び切片を、項目として含む。

【0057】次に、図7を用いて、非テキストコンテンツテーブル115のデータ構造について説明する。非テキストコンテンツテーブル115は、リンク先を持つ文書について、その文書の文書IDと、その文書の内容に関連し、その文書からリンクされている非テキストコンテンツの文書ID（以下、関連非テキストコンテンツIDという）と、その非テキストコンテンツのファイル種別を格納する。

【0058】最後に、図8を用いて、サービス種別テーブル116のデータ構造について説明する。図8に示すように、サービス種別テーブル116は、各文書について文書IDと、その各文書で提供するサービスの種別を格納する。

【0059】以下、図9から図15を用いて、文書検索装置100を構成する各部によって行われる処理について説明する。なお、ページ分類部106によって行われる処理についての説明は、上述のように省略する。

【0060】まず、収集部101は、継続してネットワークから文書を収集し、収集された文書間のリンク関係を解析し、収集及び解析結果を文書テーブル111及び

リンク関係テーブル112に格納する。人気度算出部102は、定期的に、例えば毎日、算出日の前の一定期間内に収集又は更新された文書について、人気度を算出する。なお、1日毎は例示に過ぎず、本発明を限定する趣旨ではない。以下、図9を用いて人気度を算出する処理の手順について説明する。

【0061】図9に示すように、まず、人気度算出部102は、毎日、定時に起動する。人気度を算出する人気度算出日をd1とすると、人気度算出部102は、d1からN日前、例えば150日前の日d2を算出対象開始日として決定する（ステップS11）。なお、150日は例示に過ぎない。Nは、人気度として意味のある結果を得ることができる程度に長い期間であればよい。

【0062】続いて、人気度算出部102は、収集日又は更新日が算出対象開始日d2から算出日d1までの間にあるリンク関係情報をリンク関係テーブル112から抽出する（ステップS12）。人気度を算出する対象となる文書の収集日又は更新日を一定期間内に制限することにより、作成された後、更新されずに放置されたままの文書を人気度を算出する対象から除く事が可能となる。

【0063】人気度算出部102は、抽出したリンク関係情報のうちで、同じリンク元IDを持つリンク関係情報がある場合、最新の収集日又は更新日を持つリンク関係情報を残し、その他の同じリンク元IDを持つリンク関係情報を削除する（ステップS13）。これにより、同じ文書について人気度を重複して算出することを防ぐ事が可能となる。

【0064】人気度算出部102は、抽出したリンク関係情報に基づいて、各文書の人気度を算出する（ステップS14）。より具体的には、人気度算出部102は、文書の内容を参照することなく、リンク関係及び、リンク元の文書とリンク先の文書の文書位置情報を示す文字列の類似している度合いである類似度に基づいて、各文書の人気度を算出する。以下、人気度の算出手順について説明する。

【0065】人気度を算出する際の基本的な考え方は以下の通りである。

1. 類似していない文書位置情報を持つ文書から多くリンクされている文書は、人気が高い。

【0066】例えば、一般に、同一サイト内に設けられた複数の文書はそのサイト内の他の文書にリンクされているが、それらの文書の文書位置情報は相互に類似する。従って、文書位置情報を示す文字列が相互に類似している文書からリンクされている文書の人気度は低いと推定できるからである。

【0067】2. 多くの文書からリンクされている文書ほど人気度が高い文書であり、類似していない文書位置情報を持ち人気度が高い文書からリンクされている文書の人気度は高い。

【0068】例えば、有名なディレクトリサービス等及び官公庁等は、多くの文書からリンクされているが、このような文書からリンクされている文書の方が、個人が開設するサイトやそのコンテンツのエントリページからリンクされている文書よりも人気度が高いと考えられるからである。また、多くの文書やミラーサイトを抱えるサービス(サイト)に設けられた文書等は、そのサイト内でリンクされていることが多い。1つのサイト内の文書の文書位置情報は、例えばドメインが同じ等大抵類似しているため、「文書位置情報が類似していない文書からリンクされている文書の人気度は高い」という考え方を導入すれば、サイト内で多数回リンクしあっている文書の人気度が高くなってしまふことを解消することが可能となる。

【0069】3. 文書位置情報が類似しているか否かは、サーバアドレス、パス、ファイル名の全てが異なるものが最も小さく、ミラーサイトや同一サーバ内の文書は類似度が高くなるように、文書位置情報を示す文字列から定義する。

【0070】上述の3つの考え方を導入することにより、全てのリンク関係を同等に扱わないで、リンク関係に重みを与えて扱うこととしている。より具体的には、リンク関係に重みをリンク元とリンク先文書の文書位置情報の類似度の逆数として与えることとしている。

【0071】以下、人気度を算出する手順についてより詳しく説明する。

人気度の算出対象となる文書集合を  $DOC = \{p_1, p_2, \dots, p_N\}$ 、

文書  $p$  の人気度を  $W_p$ 、文書  $p$  のリンク先の文書集合を  $Ref(p)$ 、文書  $p$  のリンク元の文書集合を  $Refed(p)$ 、文書  $p$  と文書  $q$  の文書位置情報の類似度を  $sim(p, q)$ 、相異度を  $diff(p, q) = 1/sim(p, q)$  とすると、文書  $p$  から文書  $q$  にリンクが張られているとした時、そのリンク関係の重み  $lw(p, q)$  を以下の(1)式で定義する。

【0072】

【数1】

$$lw(p, q) = \frac{diff(p, q)}{\sum_{i \in Ref(p)} diff(p, i)} = \frac{1}{sim(p, q) \sum_{i \in Ref(p)} \frac{1}{sim(p, i)}} \quad \dots (1)$$

【0073】この(1)式から分かるように、 $lw(p, q)$  は、文書  $p$  と文書  $q$  のURLの類似度  $sim(p, q)$  が低いほど、また、文書  $p$  から文書  $p$  へのリンクの数がより少ないほど大きくなる。

【0074】文書  $q$  の人気度  $W_q$  は、各文書  $p \in DOC$  に対して、 $C_q$  を定数(人気度の下限であり、文書によって異なる値を与えてもよい。)として、

【0075】

【数2】

$$W_q = C_q + \sum_{p \in Refed(q)} W_p \times lw(p, q) \quad \dots (2)$$

【0076】という(2)式に示す連立一次方程式の解として定義される。人気度算出部102は、この連立一次方程式を解くことにより、各文書の人気度を算出する。なお、このような連立一次方程式の解法については、既存のアルゴリズムが多数存在するため、説明は省略する。(1)式中の文書位置情報の類似度  $sim(p, q)$  の算出方法については後述する。(1)式及び(2)式から、上述の考え方が実現されていることを読み取ることができる。すなわち、(1)式から文書位置情報の類似度が低ければ、リンク関係の重み  $lw$  は大となる。そして、(2)式からリンク関係の重み  $lw$  が大きい文書からリンクされている文書の人気度  $W_q$  は、高くなる。つまり、類似度の低い文書位置情報を持つ文書から多くリンクされている文書の人気度は、高くなる。また、(2)式から多くの文書からリンクされている文書ほど人気度が高くなる。さらに、(2)式から人気度  $W$  が高い文書からリンクされている文書の人気度は高くなることも分かる。

【0077】次に、(1)式及び(2)式中の文書  $p$  と文書  $q$  の文書位置情報の類似度  $sim(p, q)$  について説明する。以下、文書位置情報をURLと仮定して説明するが、本発明を限定する趣旨ではない。

【0078】一般に、文書のURLは、サーバアドレス、パス、ファイル名の三種類の情報から構成される。例えば、WWW文書のURL、  
<http://www.flab.fujitsu.co.jp/hypertext/news/1999/product1.html> は、サーバアドレス([www.flab.fujitsu.co.jp](http://www.flab.fujitsu.co.jp))、パス([hypertext/news/1999](http://www.flab.fujitsu.co.jp/hypertext/news/1999))、ファイル名([product1.html](http://www.flab.fujitsu.co.jp/hypertext/news/1999/product1.html))の3種類の情報から構成される。

【0079】また、サーバアドレスは、さらに“.”により階層化されており、後ろに行くにしたがって、段々広くなる。例えば、サーバアドレスが[www.flab.fujitsu.co.jp](http://www.flab.fujitsu.co.jp)であれば、後ろから、日本(jp)、会社(co)、富士通(fujitsu)、研究所(flab)、マシン(www)という階層を表している。

【0080】本実施形態に係わるリンク関係の重みの算出方法は、以下のような考え方に基いている。

1. 往々にして、同じような文書を同一ディレクトリに入れるため、同一サーバでパスも同じ文書位置情報は内容が似ていることが多い。

2. アクセスを分散させるために設けられるミラーサイト内の文書と、オリジナルサイトの文書の文書位置情報は類似度が高い。例えば、サーバアドレス部分だけが異なり、残りのパスやファイル名は同じ場合が多い。

3. サーバアドレス、パス、ファイル名が全てことなる文書位置情報は、類似度が低い。

【0081】本実施形態では、与えられた2つの文書  $p$



及び分書qの文書位置情報の類似度を、上述のサーバアドレス、パス、ファイル名の三種類の組合せにより定義する。類似度 $\text{sim}(p, q)$ として、例えば、以下に述べるドメイン類似度 $\text{sim-domain}(p, q)$  及び融合類似度 $\text{sim-merge}(p, q)$ が考えられる。

【0082】ドメイン類似度 $\text{sim-domain}(p, q)$  は、ドメインの類似に基づいて算出される。ドメインとは、サーバアドレスの後半部分であり、会社や組織を表す。サーバアドレスが.com、.edu、.org等で終わる米国サーバの

$$\text{sim-domain}(p, q) = 1 / \alpha$$

$$= 1$$

ここで、 $\alpha$ は定数で、0より大きく1より小さい実数値を取るとする。 $\text{sim-domain}(p, q)$ の概念を導入することにより、異なるドメインを持つ文書が検索されやすくなる。言い換えると、同じドメインを持つ文書は検索されにくくなる。

【0084】 $\text{sim}(p, q)$ として、前述の三種類の情報を融合した融合類似度 $\text{sim-merge}(p, q)$ を次のように定義する。

$\text{sim-merge}(p, q) = (\text{サーバアドレスの類似度}) + (\text{パスの類似度}) + (\text{ファイル名の類似度})$

以下、右辺の各項の算出方法について説明する。

【0085】サーバアドレスの類似度は、アドレスの階層を後ろから見ていき、 $n$ レベルまで一致した場合、類似度を $1 + n$ とする。例えば、www.fujitsu.co.jp とwww.flab.fujitsu.co.jpは、相互に3レベルまで一致しているの、融合類似度は4となる。www.fujitsu.co.jp とwww.fujitsu.com は、相互に1レベルも一致していないので（一致0レベル）、融合類似度は1である。

【0086】パスの類似度は、先頭からパスの"/" で区切られた要素毎に比較し、一致したレベルまでを類似度とする。例えば、/doc/patent/index.htmlと/doc/patent/1999/2/file.htmlとは、2レベルまで一致しているので類似度は3である。

【0087】ファイル名の類似度は、ファイル名が一致する場合、類似度1とする。この $\text{sim-merge}(p, q)$ によっても、URLが似通った文書からリンクされている文書の人気度は、URLが似通っていない文書からリンクされている場合と食らえて低くなる。従って、 $\text{lw}(p, q)$ の中に $\text{sim}(p, q)$ 又は $\text{diff}(p, q)$ という概念を導入することにより、大量の文書を抱えるサーバ(サイト)や個人が単に量が多いというだけで人気度が高いことになるという問題を解消することができる。

【0088】人気度を算出した後、人気度算出部102は、各文書を人気度が高い順にソートする事により人気度順位を取得する(ステップS15)。人気度順位の時系列の変化は、増加することもあれば、減少することもある。従って、従来の算出方法による人気度の時系列の変化は増加する一方であったという問題は、人気度の代

場合はサーバアドレスの後ろから2つめまで、サーバアドレスが.jp、.fr 等で終わる他国のサーバの場合はサーバアドレスの後ろから3つめまでがドメインに相当する。例えば、www.fujitsu.com のドメインはfujitsu.comであり、www.flab.fujitsu.co.jpのドメインはfujitsu.co.jp である。

【0083】文書pと文書qのドメイン類似度は以下の(3)式により定義される。

$$(p, q \text{ が同一ドメインの場合})$$

$$(p, q \text{ が異なるドメインの場合})$$

$$\dots (3)$$

わりに人気度順位の時系列の変化に注目する事によっても解決する事が可能となる。最後に、人気度算出部102は、算出した人気度及び人気度順位を、各文書の文書ID及び人気度算出日とともに人気度テーブル113に格納し(ステップS16)、処理を終了する。

【0089】例えば、ユーザに文書を検索した結果を提供する際に、上述のように算出した人気度に基づいて各文書は、ソート又はランキングされることとしてもよい。また、ある文書に関する情報を提供する際に、その文書の人気度もユーザに提供することとしても良い(後述)。

【0090】以下、図10を用いて、人気度の算出における本発明の特徴について説明する。図10(a)は、従来の算出方法によって算出した人気度の時間的変化を示す図である。図10(a)において、横軸は時間、縦軸は人気度を示す。ウェブにおいて一度作成された文書を作者が削除したり、更新したりすることはあまりないため、従来技術のように単純に文書へリンクしている他の文書の数(被リンク数)に基づいてその文書の人気度を算出すると、人気度が減ることはなく、図10(a)に示すように増加する一方となる。

【0091】図10(b)は、本発明に係わる算出方法によって算出した人気度の時間的変化を示す図である。図10(b)においても、横軸は時間、縦軸は人気度を示す。本発明によれば、算出対象開始日から人気度算出日間での間の一定期間内に収集又は更新された文書について人気度を算出するため、従来のように、一度作成された後、長期間放置されたままの文書は人気度を算出する対象とならない。従って、例えば、長期間放置されたままの文書をリンク元とする文書の人気度は、従来よりも低く算出されることとなる。これにより、従来、人気度が増加する一方であったという問題を解決する。

【0092】また、例えば、ウェブで公開されたばかりのサイトのトップページは、そのサイト内の文書等から多くリンクされているため、そのトップページの人気度は当初高く算出されるが、その後サイト内の文書が更新されずに放置されると、そのトップページの人気度は低下し、人気度の高い状態は一過性のものとなる。

【0093】図10(b)に示す文書の人気度は、当初、人気度が急に上昇しているが、ある程度の時間が経過した後、人気度は減少に転じ、以後減少しつづけている。このことから、この文書の流行は一過性に終わった事がわかる。

【0094】図10(c)は、本発明に係わる算出方法によって算出した人気度に基づく人気度順位の時間的変化を示す図である。図10(c)において、横軸は時間、縦軸は人気度順位を示す。人気度順位は、人気度を算出する対象となる文書全体から見たその文書の相対的な人気度を示す情報であるため、その性質から従来の算出方法によって人気度を算出した場合でも、増加しつづける事はあまり考えられない。従って、人気度順位の時間的変化に基づいて文書の人気度を判断することによっても、従来、人気度が増加する一方であったという問題を解決することができる。

【0095】また、本発明に係わる算出方法によって算出した人気度に基づく人気度順位の時間的変化によれば、その文書が、人気度を算出する対象となる文書全体から見て平均的な順位の推移を示す場合、図10(b)のグラフに示すように、人気度順位は、時間が経過してもほぼ一定に推移する。また、その文書の人気度が増加している場合、人気度順位も人気度の増加にあわせて上昇する。他方、その文書の人気度が減少している場合、人気度順位は、人気度の減少にあわせて下降する。一般に、文書の人気は、当初増加期から始まり、安定期を経て減少期に至る。この場合、図10(c)に示すように、人気度順位は、増加期には上昇し、安定期にいたるとはほぼ一定になり、減少期には下降するため、人気度順位の時間的変化は山形になる。

【0096】次に、図11を用いて、人気変化度を算出する処理の手順について説明する。人気算出部102が人気度を算出すると、人気度遷移算出部103は、一定期間内に算出された人気度を人気度テーブル113から取得し、人気度の時間的変化である人気変化度を算出する。

【0097】まず、人気度遷移算出部103は、人気度算出日d1からM日、例えば14日前の日d3を算出対象開始日として決定する(ステップS21)。なお、14日は例示に過ぎない。Mは、あまり長く取ると、短期的な人気度の変動を把握する事ができなくなるため、数週間程度にすることが望ましい。

【0098】続いて、人気度遷移算出部103は、各文書について算出対象開始日d3から人気度算出日d1の間に算出された人気度又は人気度順位を人気度テーブル113から取得する(ステップS22)。人気度遷移算出部103は、各文書ごとに、人気度又は人気度順位の時間に対する線形回帰式を算出し、その線形回帰式の回帰係数及び切片bを得る(ステップS23)。人気度に基づいて線形回帰式を算出した場合、回帰係数aが人気

変化度に相当し、人気度順位に基づいて線形回帰式を算出した場合、回帰係数aを切片bで除算した値、 $a/b$ が人気変化度に相当する。

【0099】以下、線形回帰式の算出方法について詳しく説明する。日付d3からd1までの(d3, d3+1, ..., d1)のそれぞれの日付における人気度又は人気度順位の値を( $w_0, w_1, \dots, w_{M-1}$ )とすると、線形回帰式

$$r = a(d1 - d3) + b$$

は、最小二乗法によって算出される。ここで、aは回帰係数であり、以下の式で算出される。

【0100】

$$a = (M \times Iw - I \times W) / (M \times I2 - I^2)$$

また、bは切片であり、以下の式で算出される。

$$b = (I \times Iw - W \times I2) / (I2 - M \times I2)$$

ここで、Iw、W、I及びI2は、それぞれ以下の式で算出される。

【0101】

【数3】

$$Iw = \sum_{i=0}^{M-1} i \cdot w_i$$

【0102】

【数4】

$$W = \sum_{i=0}^{M-1} w_i$$

【0103】

【数5】

$$I = \sum_{i=0}^{M-1} i = \frac{M(M-1)}{2}$$

【0104】

【数6】

$$I2 = \sum_{i=0}^{M-1} i^2 = \frac{M(M-1)(2M-1)}{6}$$

【0105】最後に、人気度遷移算出部103は、算出した各文書の回帰係数a及び切片bを、文書IDとともに人気度変化テーブル114に格納し(ステップS24)、処理を終了する。

【0106】人気度に基づいて線形回帰式を算出した場合、線形回帰式の回帰係数aが正であれば人気度が上昇中であり、その絶対値が大きいほど、人気度が上昇する速度が速いことを示す。また、切片bが比較的高い値以上である場合、人気度が高い水準で安定している事を示し、切片bが比較的低い値以下である場合、人気度が低い水準で安定していることを示す。

【0107】一方、人気度順位に基づいて線形回帰式を算出した場合、回帰係数aが負であれば人気度が上昇中であり、その絶対値が大きいほど、人気度が上昇する速度が速いことを示す。また、切片bが比較的低い値以下である場合、人気度が高い水準で安定している事を示し、切片bが比較的高い値以上である場合、人気度が低

い水準で安定していることを示す。

【0108】各文書の人気変化度は、その文書についての情報をユーザに提供する際に、その文書の文書位置情報、タイトル及び内容を示す情報とともに、ユーザに提供される。提供される際、人気変化度は、数値としてではなく、人気度の変化の方向及び度合いを図示するアイコンを用いて提供される事としても良い(後述)。

【0109】次に、図12を用いて各文書の内容に関連する関連非テキストコンテンツを判定する処理について説明する。文書中には、テキストコンテンツ以外にも、画像、音声等の非テキストコンテンツが含まれる事が多い。そして、文書中に含まれる非テキストコンテンツの中には、バナー広告等、文書の内容に関係ない非テキストコンテンツもある。関連非テキストコンテンツ判定部104は、リンク関係に基づいて、文書中に含まれる非テキストコンテンツが文書の内容に関連するか否かを判定する。

【0110】そのために、まず、非テキストコンテンツ判定部104は、リンク関係テーブル112を参照し、リンク先IDが格納されているリンク関係情報を抽出する。なお、抽出されたリンク関係情報のうち、同じリンク元IDを持つリンク関係情報ある場合、最新の収集日又は更新日を持つリンク関係情報のみを採用し、その他は削除する。同じ文書について同じ処理を行う事を防ぐためである。

【0111】以後、抽出されたリンク関係情報に含まれるリンク元IDによって特定されるリンク元文書Sからなる文書集合をリンク元文書集合とする。抽出されたリンク関係情報に含まれるリンク先IDによって特定される文書(つまり、リンク先文書)は、判定対象文書Cという。

【0112】ステップS31からステップS40までの手順は、各リンク元文書Sに含まれる判定対象文書Cそれぞれについて行う。まず、非テキストコンテンツ判定部104は、各リンク元文書Sから判定対象文書Cへリンクする部分の近辺に存在するリンク文字列Aを抽出する(ステップS31)。

【0113】例えば、HTMLを用いた文書の場合、非テキストコンテンツ判定部104は、アンカータグ(<a>)の前後100バイトをリンク文字列Aとして抽出することとしても良い。続いて、関連非テキストコンテンツ判定部104は、そのリンク文字列Aが特定の文字列であるか否かを判定する(ステップS32)。

【0114】特定の文字列とは、例えば、「MPEG」、「動画」、「ストリーミング」、「video」、「audio」及び「mp3」や、動画等のフォーマット名など、非テキストフォーマットを示す文字列である。これらの特定の文字列を定義するテーブルは、予め、文書検索装置100に備えられているものとする(不図示)。

【0115】関連非テキストコンテンツ判定部104は、そのリンク文字列Aが特定の文字列であると判定した場合(ステップS32:Yes)、その判定対象文書Cをリンク元文書Sの内容に関連する非テキストコンテンツとして判定し、ステップS40に進む。関連非テキストコンテンツ判定部104は、その判定対象文書Cの種別及びリンク元文書Sの文書IDとともに、その判定対象文書Cの文書IDを関連非テキストコンテンツIDとして、非テキストコンテンツテーブル115に格納し、その判定対象文書Cについての処理を終了する。

【0116】関連非テキストコンテンツ判定部104は、そのリンク文字列Aが特定の文字列でないと判定した場合(ステップS32:No)、更に、判定対象文書Cの文書位置情報に含まれる判定対象文書Cのファイル名の拡張子が、特定の拡張子であるか否かを判定する(ステップS33)。

【0117】現在のウェブでは、特定の拡張子として、例えば、以下のようなものが考えられる。なお、各拡張子についての説明は、当業者に自明であるため省略する。なお、この例示は、本発明を限定する趣旨ではない。

・音楽系のコンテンツの場合

mp3、wma、wav

・動画系のコンテンツの場合

ram、rm、rv、rmm、wmv、avi、asx、qt、mov、mpeg、mpg、fla、swf

・画像系のコンテンツの場合

jpg、jpeg

関連非テキストコンテンツ判定部104は、このような拡張子によっても、判定対象文書Cが非テキストコンテンツであるか否かを判定する事ができる。これらの特定の拡張子を定義するテーブルは、予め、文書検索装置100に備えられているものとする(不図示)。関連非テキストコンテンツ判定部104は、判定対象文書Cの文書位置情報に含まれるファイル名の拡張子が特定の拡張子でないと判定した場合(ステップS33:No)、判定対象文書Cは非テキストコンテンツでないと、その文書についての処理を終了する。

【0118】関連非テキストコンテンツ判定部104は、判定対象文書Cのファイル名の拡張子が特定の拡張子であると判定した場合(ステップS33:Yes)、更にその判定対象文書Cにリンクとして使用されているか否かを判定する。この判定は、例えば、HTMLの場合タグに基づいて行う事ができる。判定対象文書Cがリンクとして使用されているとは、例えば、バナー広告画像のように、その文書を選択(クリック、或いはタッチ等)することによって他の文書を閲覧することができることを意味する。

【0119】例えば、HTMLで記述された文書中で判定対象文書C(例の場合、画像)がリンクとして使用さ

れている場合、以下のように表記されることが多い。なお、この例示は、本発明を限定する趣旨ではない。

【0120】<a href="判定対象文書Cのリンク先文書の文書位置情報"></a>;

関連非テキストコンテンツ判定部104は、判定対象文書C及びリンク元文書Sの文書IDを用いて文書テーブル111を参照し、両者の文書位置情報を取得する。そして、関連非テキストコンテンツ判定部104は、判定対象文書Cの文書位置情報及びリンク元文書Sの文書位置情報に基づいて、判定対象文書Cが格納されているサイトが、リンク元文書Sが格納されているサイトと同じであるか否かを判定する(ステップS35)。

【0121】より具体的には、文書位置情報が例えばURLである場合、関連非テキスト判定部104は、判定対象文書CのURLとリンク元文書SのURLのサーバアドレス又はドメインに基づいて、判定対象文書Cが格納されているサイトとリンク元文書Sが格納されているサイトが同じであるか否かを判定する。

【0122】判定対象文書Cが格納されているサイトとリンク元文書Sが格納されているサイトが同じであると判定する場合(ステップS35: Yes)、判定対象文書Cは、リンク元文書Sの内容に関連する文書であると推測できるため、ステップS37に進む(後述)。これは、判定対象文書Cがリンク元文書Sの内容と関連している場合、判定対象文書Cは、リンク元文書Sが格納されているサイトと同じサイトに格納されている事が多いからである。

【0123】一方、判定対象文書Cが格納されているサイトとリンク元文書Sが格納されているサイトが異なると判定する場合(ステップS35: No)、関連非テキストコンテンツ判定部104は、更に、判定対象文書Cの文書位置情報及び判定対象文書Cのリンク先の文書の文書位置情報に基づいて、判定対象文書Cのリンク先となっている文書が格納されているサイトが、リンク元文書Sが格納されているサイトと同じであるか否かを判定する(ステップS36)。なお、判定対象文書Cのリンク先の文書の文書位置情報は、上記例のようにリンクを埋め込むタグ付近に記載されていることが多い。

【0124】判定対象文書Cのリンク先となっている文書が格納されているサイトが、リンク元文書Sが格納されているサイトと同じであると判定する場合(ステップS36: Yes)、ステップS37に進む。判定対象文書Cのリンク先となっている文書がリンク元文書Sの内容と関連していると推測されるため、判定対象文書Cもリンク元文書Sの内容と関連していると推測できるからである。

【0125】一方、判定対象文書Cのリンク先となっている文書が格納されているサイトが、リンク元文書Sが格納されているサイトと異なると判定した場合(ステッ

プS36: No)、関連非テキストコンテンツ判定部104は、判定対象文書Cは、バナー広告等、リンク元文書Sの内容と関連しない文書であると推定し、その判定対象文書についての処理を終了する。

【0126】ステップS37において、関連非テキストコンテンツ判定部104は、判定対象文書Cがリンク元文書S内で所定回数、例えば、3回以上使用されているか否かを判定する。なお、3回は、例示に過ぎず、本発明を限定する趣旨ではない。判定対象文書Cがリンク元文書S内で3回以上使用されているかと判定した場合(ステップS37: Yes)、関連非テキストコンテンツ判定部104は、その判定対象文書Cをリンク元文書Sの内容に関連しないと判定し、その判定対象文書Cについての処理を終了する。そうでない場合、ステップS38に進む。

【0127】例えば、判定対象文書Cが、リストのブリット等のフォーマット、或いは文書作成用の素材である場合、1つの文書内で複数回使用される可能性が高い。このような文書は、リンク元文書Sの内容とは関連がないと考えられるため、関連非テキストコンテンツとして扱わないこととする。

【0128】ステップS37でNoであった場合、関連非テキストコンテンツ判定部104は、更に、リンク元文書Sのリンク関係情報に含まれるリンク先IDに基づいて文書テーブル111からリンク元文書Sのリンク先文書のファイル名を取得し、リンク元文書Sが、判定対象文書Cと類似したファイル名を持つ他のリンク先文書を有するか否かを判定する(ステップS38)。

【0129】判定対象文書Cと類似したファイル名を持つ他のリンク先文書をリンク元文書Sが有しないと判定した場合(ステップS38: No)、ステップS40に進み、関連非テキストコンテンツ判定部104は、上述のようにして、その判定対象文書Cを非テキストコンテンツテーブル115に登録する。

【0130】判定対象文書Cと類似したファイル名を持つ他のリンク先文書をリンク元文書Sが有すると判定した場合(ステップS38: Yes)、関連非テキストコンテンツ判定部104は、判定対象文書Cが、判定対象文書Cと、それと類似するファイル名を持つリンク先文書の中で、辞書順で最も若いファイル名を持つのか否かを判定する(ステップS39)。辞書順とは、例えば、アルファベットの先の順或いは、数字では小さい順ということを意味する。

【0131】関連非テキストコンテンツ判定部104は、判定対象文書Cが、辞書順で最も若いファイル名を持つと判定した場合(ステップS39: Yes)、ステップS40に進み、判定対象文書Cを非テキストコンテンツテーブル115に登録し、その文書についての処理を終了する。そうでない場合(ステップS39: No)、ステップS40を行わないで、その文書について

の処理を終了する。

【0132】例えば、リンク元文書Sがアルバムのように画像を一覧表示する内容の文書である場合、これらの全てをリンク元文書Sの内容に関連する文書として扱おうと、関連する文書が多くなり、かえって利用者に検索結果を提供する際に煩雑となってしまうことが考えられる。しかし、このような場合、例えば、pict01.jpg、pict02.jpg、pict03.jpg、・・・のように、数値部分を除いた残りの部分は互いに同一である事が多い。従って、互いに類似したファイル名を持つリンク先文書がある場合、辞書順に最も若いファイル名を持つ文書のみを関連非テキストコンテンツとして登録することにより、このような煩雑さを避けることが可能となる。

【0133】上述のようにして、ある判定対象文書Cについての処理を終了した後、関連非テキストコンテンツ判定部104は、リンク元文書Sのリンク関係情報を参照し、先に取り出したリンク元文書Sに他の未判定のリンク先文書があるか否かを判定する。未判定のリンク先文書がある場合、関連非テキストコンテンツ判定部104は、その未判定のリンク先文書を新たな判定対象文書Cとし、その文書についてステップS31以降の処理を行う。

【0134】また、そのリンク元文書Sに他の未判定のリンク先文書が含まれていない場合は、関連非テキストコンテンツ判定部104は、他の未処理のリンク元文書Sをリンク元文書集合から取り出して、そのリンク元文書Sのリンク先文書Cについて同様の処理を行う（不図示）。また、全てのリンク元文書Sについて処理を行った場合、関連非テキストコンテンツ判定処理を終了する。

【0135】各文書についての情報をユーザに提供する際に、その文書の文書位置情報、タイトル及び内容を示す情報とともに、上記判定結果に基づいてその文書からリンクされている関連非テキストコンテンツの種別を示す情報、例えばアイコンをユーザに提供することとしても良い。これにより、ユーザは、その文書のリンク先にどのような関連非テキストコンテンツがあるのか、その文書を実際に閲覧（ブラウズ）することなく知ることができる。また、さらに、上述の関連非テキストコンテンツの種別を示すアイコンに、その関連非コンテンツへのリンクを埋め込む事により、ユーザがアイコンを選択（クリック、或いはタッチ等）した場合に、その関連非テキストコンテンツをユーザの画面に表示又は再生等させることとしても良い（後述）。

【0136】次に、図13を用いて文書のサービス種別を判定する処理の手順について説明する。文書において、様々なサービスがその文書の閲覧者に提供されることが多い。サービス種別判定部105は、文書中で用いられているフォームタグに基づいて、その文書で提供されているサービスの種別を判定する。以下の説明に

おいて、検索、ショップ及び申込（登録）の3つのサービス種別を判定している。

【0137】ここで、検索サービスとは、ユーザ（又は閲覧者等）が入力されたキーワードに基づいて何かを探すサービスをいう。ショップサービスとは、ユーザに商品を販売するサービスをいう。申込（登録）サービスとは、ユーザから氏名や住所等を受け付け、ユーザから会員や懸賞の申込又は登録を受け付けるサービスをいう。なお、これらの3つのサービスは、例示であり、本発明を限定する趣旨ではない。サービス種別を判定する処理に、さらに多くの手順を追加することによって、更に詳しくサービス種別を判定することが可能となる。

【0138】まず、サービス種別判定部105は、収集済みの文書のうちテキストが含まれる文書を抽出する（不図示）。テキストが含まれているか否かは、例えば各文書のファイル名の拡張子に基づいて判定する事にしても良い。以下の処理は、抽出された各文書について行われる。

【0139】続いて、サービス種別判定部105は、文書にフォームタグが含まれるか否かを判定する（ステップS41）。文書にフォームタグが含まれない場合（ステップS41：No）、その文書はサービスを提供していないと推測されるため、その文書についての処理を終了する。

【0140】文書にフォームタグが含まれる場合（ステップS41：Yes）、サービス種別判定部105は、更に、その文書に含まれるボタンに「購入」又は「買う」等の文字があるか否かを判定する（ステップS42）。

【0141】例えば、HTMLで記述された文書の場合、ボタンは以下のように表記されることが多い。  
 <INPUT TYPE="submit" VALUE="ボタンに表示する文字">;  
 ボタンに「購入」、「purchase」又は「買う」等の文字がある場合（ステップS42：Yes）、サービス種別判定部105は、その文書で提供されるサービスの種別を「ショップ（販売店）」であると判定し（ステップS43）、ステップS48に進む。サービス種別判定部105は、その文書の文書IDとともに判定したサービス種別「ショップ」をサービス種別テーブル116に格納する事により、その文書のサービス種別を「ショップ」として登録する（ステップS48）。

【0142】ボタンに「購入」又は「買う」等の文字がない場合（ステップS42：No）、サービス種別判定部105は、更に、その文書にユーザの入力エリアが含まれるか否かを判定する（ステップS44）。ユーザの入力エリアが含まれない場合（ステップS44：No）、その文書でサービスは提供されていないと推測し、その文書についての処理を終了する。その文書にユーザの入力エリアが含まれる場合（ステップS44：Yes）、サービス種別判定部105は、更に、その文書に含まれ

るボタンに「検索」又は「search」等の文字があるか否かを判定する(ステップS45)。

【0143】ボタンに「検索」又は「search」等の文字がある場合(ステップS45:Yes)、サービス種別判定部105は、その文書が提供するサービスの種別を「検索」とであると判定し(ステップS46)、ステップS48に進む。ステップS48において、サービス種別判定部105は、上述のようにしてその文書が提供するサービスを登録する。

【0144】ボタンに「検索」又は「search」等の文字がない場合(ステップS45:No)、サービス種別判定部105は、その文書が提供するサービスの種別を「申込」とであると判定し(ステップS47)、ステップS48に進む。

【0145】このように、サービス種別判定部105は、文書の内容を見ることなく、フォームタグに基づいて、その文書で提供されているサービスの種別を判定することができる。

【0146】なお、サービス種別を判定する処理には、様々な変形が考えられる。例えば、ステップS45とステップS46の間で以下の処理を行う事としてもよい。まず、ステップS45の後、サービス種別判定部105は、更に、ISBN(International Standard Book Number:国際標準図書番号)の入力欄があるか否かを判定し、ISBNの入力欄が含まれる場合、その文書が提供するサービスの種別を「書店」として判定してステップS48に進む。ISBNの入力欄が含まれない場合、ステップS46に進む。これにより、文書が提供しているサービスを更に詳しく判定する事が可能となる。

【0147】各文書についての情報をユーザに提供する際に、その文書の文書位置情報、タイトル及び内容を示す情報とともに、上記判定結果に基づいて、その文書が提供するサービスの種別を示す情報、例えばアイコンをユーザに提供することとしても良い。これにより、ユーザは、その文書が提供しているサービスの種別をその文書を実際に関連(ブラウズ)することなく知ることができる。また、上記判定において判定されたサービス種別が、各ページを分類する際に使用する事ができる。

【0148】ページ分類部106は、各文書中の語句に基づいて、その文書の内容を判定し、判定結果に基づいて各文書を分類する。文書の内容を示す語句として、例えば、「Java(登録商標)」、「テーマパーク」等が考えられる。なお、この例示は本発明を限定する趣旨ではない。このページ分類部による各文書の分類方法は従来技術と同じであるため、詳しい説明は省略する。なお、ページ分類部106は、各文書を分類する際に、例えば、サービス種別判定部105によって判定された各文書で提供されるサービス種別を利用することとしても良い。

【0149】検索サービス部107は、文書検索装置1

00のユーザからの指示に基づいて文書を検索し、適宜上述の人気度算出部102及び人気度遷移算出部103等の処理結果とともに検索結果をそのユーザに対して提供する。より具体的は、検索サービス部107は、ユーザの端末に処理結果とともに検索結果を表示させる。以下、検索サービス部107が行う処理について、ユーザの端末に表示される画面を適宜参照しながら説明する。

【0150】検索サービス部107は、検索の結果得られた文書に関する情報を、さまざまな形式でユーザに提供する。まず、ユーザがキーワード等を入力し、そのキーワード等に基づいて検索した結果をユーザに提供する場合について説明する。

【0151】まず、検索サービス部107は、ユーザが入力したキーワード等に基づいて、文書を検索し、検索された文書について、以下の情報を各テーブルから取得する。

【0152】・最新の人気度及び人気度順位を人気度テーブル113から取得する。

・最新の人気度及び人気度順位のそれぞれに基づく回帰係数 $a$ (傾き)及び切片 $b$ を人気度変化テーブル114から取得する。

【0153】・関連非テキストコンテンツの文書IDを非テキストコンテンツテーブル115から取得する。

・サービス種別をサービス種別テーブル116から取得する。

【0154】続いて、検索サービス部107は、取得した回帰係数 $a$ 及び切片 $b$ に基づいて、人気度の変化の方向と速度を図示する人気度推移アイコンを作成する。人気度推移アイコンは、具体的には、矢印を図示するアイコンであり、人気度の変化の方向と速度を矢印の向きと傾きで示す。検索サービス部107は、人気度推移アイコンとして、例えば、以下の6種を作成する。なお、この例示は、本発明を限定する趣旨ではない。

【0155】急上昇アイコン：人気度が急激に上昇している事を示す。急上昇アイコンは、角度が急な右肩上がりの矢印を図示する。

上昇アイコン：人気度が上昇している事を示す。上昇アイコンは、右肩上がりの矢印を図示し、その角度は、急上昇アイコンよりも水平に近い。

【0156】下降アイコン：人気度が下降している事を示す。下降アイコンは、右肩下がりの矢印を図示し、その角度は、急下降アイコンよりも水平に近い。

急下降アイコン：人気度が急激に下降している事を示す。急上昇アイコンは、角度が急な右肩下がりの矢印を図示する。

【0157】安定アイコン：右向きの水平の矢印を図示する。後述の高値安定と低値安定の場合とで色を変えることとしてもよい。

無印アイコン：矢印がないアイコンである。その他の状態を示す。

【0158】人気度推移アイコンの作成方法の例として、以下の2つを挙げる。

(例1) 人気度変化を人気度(10000までの自然数。大きいほど人気度が高い)を元に計算した場合  
検索サービス部107は、以下のようにして回帰係数a及び切片bに基づいて各文書に付すべきアイコンを判定する。

【0159】急上昇アイコン：その文書のaが50以上の場合

上昇アイコン：その文書のaが30以上の場合

下降アイコン：その文書のaが-30以下の場合

急下降アイコン：その文書のaが-50以下の場合

高値安定アイコン：その文書のbが8000以上の場合

低値安定アイコン：その文書のbが3000以下の場合

無印アイコン：その他の場合

(例2) 人気度変化を人気度順位(1から総文書数までの自然数。小さいほど人気度順位がよい)で計算した場合

検索サービス部107は、以下のようにして各文書に付すべきアイコンを判定する。

【0160】急上昇アイコン：その文書のa/bが-0.1以下(10%以上増加)の場合

上昇アイコン：その文書のa/bが-0.05以下(5%以上増加)の場合

下降アイコン：その文書のa/bが0.05以上(5%以上減少)の場合

急下降アイコン：その文書のa/bが0.1以上(10%以上減少)の場合

高値安定：その文書のbが1000以下の場合

低値安定：その文書のbが100000以上の場合

無印：その他の場合

続いて、検索サービス部107は、関連非テキストコンテンツが登録されていた文書について、関連非テキストコンテンツの種類を図示する関連メディアアイコンを作成し、その関連メディアアイコンに関連非テキストコンテンツへのリンクを埋め込む。これにより、関連メディアアイコンをユーザが選択すると、その関連非テキストコンテンツのリンク元文書を閲覧することなく関連非テキストコンテンツを閲覧、再生等させることが可能となる。

【0161】関連メディアアイコンは、例えば、関連非テキストコンテンツの種類を表示する。より具体的には、関連非テキストコンテンツが「jpg」形式である場合、関連メディアアイコンは、「jpg」という文字列を表記する。或いは、関連メディアアイコンは、画像を示すように、カメラを図示する事としても良い。なお、文書に複数の関連非テキストコンテンツが登録されている場合、各関連非テキストコンテンツについてこの処理を行う。

【0162】さらに、検索サービス部107は、サービ

ス種別が登録されていた文書について、サービス種別の種類を図示するサービス内容アイコンを作成する。サービス内容アイコンは、例えば、サービスの種別を表示するアイコンである。より具体的には、サービス種別がショップである場合、サービス内容アイコンは、「ショップ」という文字列を表記する。或いは、サービス内容アイコンは、ショップを図示する事としても良い。

【0163】最後に、検索サービス部は、検索の結果得られた各文書を人気度順位に基づいてソートし、ソートした順に、各文書のタイトル、文書の内容を示す情報、文書の文書位置情報、人気度推移アイコン、関連メディアアイコン及びサービス内容アイコンを画面に設定する。これにより、図14に示すような、検索結果の表示画面が作成される。

【0164】図14に示す検索結果の表示画面において、各文書は、最新の人気度の順、つまり静的な人気度の順に並べられる。ユーザは、各文書の人気度がどのように変化した結果、この順位になったのか、人気度推移アイコンによって知ることができる。さらに、ユーザは、関連メディアアイコンによって、各文書はどのような非テキスト文書にリンクしているのか知ることができ、さらに、関連メディアアイコンを選択(クリック、或いはタッチ等)することにより、関連非テキストコンテンツを再生、又は閲覧等する事が可能である。従って、ユーザは、その文書を閲覧することなく、その文書からどのような非テキストコンテンツにリンクしているのかを知ることが可能となる。

【0165】また、さらに、ユーザは、サービス内容アイコンによって、各文書はどのようなサービスを提供しているのか知ることができる。図14において、ユーザが人気度推移アイコンを選択(クリック、或いはタッチ等)すると、検索サービス部107は、人気度推移アイコンが選択された文書について、過去一定期間内、例えば各数ヶ月内に算出された人気度又は人気度順位を人気度テーブル113から取得し、人気度が算出された日付に対する人気度又は人気度順位のグラフを作成し、画面に設定する。

【0166】図15(a)に、人気度が算出された日付に対する人気度順位のグラフが設定された人気度推移画面の一例を示す。図15(a)において、横軸が日付、縦軸が人気度順位を示す。また、グラフ中において数字は上下に記載されているが、上段の数字は人気度順位を示し、下段の数字は人気度が算出された日付を示す。このグラフは、当該文書の人気度が、この数ヶ月どのように推移したのかを示したものであり、人気度変化テーブルを視覚化したものに相当する。図15(a)に示すように、URL: www.aaaによって特定される文書の人気度順位は、3月に急上昇した後、5月以降ほぼ安定して推移している事が分かる。

【0167】図15(a)において、グラフ中の一部が

選択されると、検索サービス部107は、その選択された付近の適当な期間内の日付を収集日又は更新日とし、その文書の文書IDをリンク先IDとするリンク関係情報をリンク関係テーブル112から取得する。そして、検索サービス部107は、取得したリンク関係情報に基づいて、その一定期間内にその文書をリンク先としていた文書の一覧を作成し、画面に設定する。

【0168】図15(b)にある期間内において、URL: www.aaaで特定される文書をリンク先としていた文書、つまり、URL: www.aaaで特定される文書のリンク元文書の一覧を示す画面の一例を示す。図15(b)によって、ユーザは、その時期に、その文書がどのような文書からリンクされているのか知ることができる。例えば、ユーザが、URL: www.aaaで特定される文書のサイトマスターである場合、ユーザは、今後のサイトのメンテナンスにこの情報を応用する事が可能となる。

【0169】また、更に、ユーザは、予めある文書の文書位置情報及び人気度の閾値を検索サービス部107に登録しておき、検索サービス部107は、その文書の人気度が閾値以上又は閾値以下になった場合に、そのユーザに通知する事としてもよい。この場合も、ユーザは、その文書の人気度の変化を自動的に知ることができるため、ユーザは、今後のサイトのメンテナンス等にこの情報を応用する事が可能となる。

【0170】また、本発明の文書検索装置は、一般的な検索以外のその他、様々な用途に利用可能である。例えば、文書検索装置100を、業界分析ツールとして利用することもできる。文書検索装置100を利用して特定業界の人気度推移を表示し、ユーザはこの人気度推移をマーケティングの助けにすることができる。そのために、利用者は、まず、知りたい業界の企業トップページ(文書)の文書位置情報の一覧(例えばURL集)を作成する。

【0171】続いて、文書検索装置100は、文書位置情報の一覧に含まれる各文書の最新の人気度を人気度テーブル113から取得し、取得した人気度が高い順に各文書を一覧表示した人気度リストを設定する。この人気度リストは現在の業界ランキングを意味する。

【0172】図16(a)に、人気度リストの一例を示す。図16(a)の下端に「過去1ヶ月」及び「過去1年」と表示されたボタンが設定されている。このボタンが押下されると、文書検索装置は、さらに、過去1ヶ月間又は過去1年間に算出された文書位置情報一覧に含まれる各文書の人気度を人気度テーブル111から取得し、人気度を算出した日付に対する人気度の推移を示すグラフを作成し、画面に設定する。なお、人気度の代わりに人気度順位を用いても良いことはいうまでもない。

【0173】図16(b)に、過去1年の各文書の人気度の推移を示すグラフの一例を示す。図16(b)は、

図16(a)に示すリスト内の各文書の過去1年の人気度の推移を示し、図16(a)において「過去1年」と表記されたボタンが押下された場合に、ユーザの端末に表示される。図16(b)において、横軸は、人気度が算出された日付を、縦軸は人気度を示す。図16(b)に示すように、URL: bbb.co.jpを持つ文書の人気度が過去1年で急上昇している事が分かる。

【0174】また、例えば、文書検索装置100を、地域情報検索システムとして利用する事も可能である。そのために、まず、ページ分類部106は、例えば、都道府県、市町村等のような地域を示す階層的なカテゴリを作成し、そのカテゴリに従って各文書を分類する。ユーザは、階層的なカテゴリを辿って、求める文書とその人気度、人気度推移、参照メディア、ページで提供するサービスにアクセスさせることができる。

【0175】図17に、地域情報検索システムの画面の一例を示す。図17(a)に、カテゴリ「東京都」に関する文書を一覧表示する画面の一例を示す。図17(a)において、画面の上段に選択された地域「東京都」が表示され、中段に東京都内の各区が表示され、下段に「東京都」に分類された各文書に関する情報が表示されている。画面の下段は、図14に示す検索結果の表示画面と同様であるため、図17において省略している。図17(a)の画面の上段においてユーザが「港区」を選択すると、カテゴリ「港区」に関する文書を一覧表示する画面に遷移する。

【0176】図17(b)に、カテゴリ「東京都-港区」に関する文書を一覧表示する画面の一例を示す。図17(b)において、画面の上段に選択された地域「港区」が表示され、画面の中段に港区内の町名が表示され、画面の下段に「東京都-港区」に分類された各文書に関する情報が表示されている。画面の下段は、図14に示す検索結果の表示画面と同様である。図17(b)の画面の上段においてユーザが更に「六本木」を選択すると、カテゴリ「東京都-港区-六本木」に関する文書を一覧表示する画面に遷移する。

【0177】図17(c)に、カテゴリ「東京-港区-六本木」に関する文書を一覧表示する画面の一例を示す。図17(c)において、画面の上段に選択された地域「六本木」が表示され、画面の中段にその他のカテゴリが表示され、画面の下段に「東京-港区-六本木」に分類された文書に関する情報が表示されている。

【0178】本実施形態において説明した文書検索装置100及びユーザの端末等は、図18に示すようなコンピュータ(情報処理装置)を用いて構成することもできる。図18のコンピュータ200は、CPU201、メモリ202、入力装置203、出力装置204、外部記憶装置205、媒体駆動装置206、及びネットワーク接続装置207を備え、それらはバス208により互いに接続されている。



【0179】メモリ202は、例えば、ROM (Read Only Memory)、RAM (Random Access Memory) 等を含み、処理に用いられるプログラムとデータを格納する。CPU201は、メモリ202を利用してプログラムを実行することにより、必要な処理を行う。

【0180】コンピュータ200に文書検索装置100に相当する機能を実現させる場合、図1に示す文書検索装置100を構成する収集部101、人気度算出部102、人気度遷移算出部103、関連非テキストコンテンツ判定部104、サービス種別判定部105、ページ分類部106及び検索サービス部107は、各部によって行われる処理を示すプログラムとして実現され、それぞれメモリ202の特定のプログラムコードセグメントに格納される。なお、上述の各部によって行われる処理は、各フローチャートにおいて説明されている。

【0181】入力装置203は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置204は、例えば、ディスプレイやプリンタ等であり、コンピュータ200の利用者への問い合わせ、処理結果等の出力に用いられる。

【0182】外部記憶装置205は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク装置等である。この外部記憶装置205に上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ202にロードして使用することもできる。

【0183】媒体駆動装置206は、可搬出記録媒体209を駆動し、その記録内容にアクセスする。可搬出記録媒体209としては、メモリカード、メモリスティック、フレキシブルディスク、CD-ROM (Compact Disc Read Only Memory)、光ディスク、光磁気ディスク、DVD (Digital Versatile Disk) 等、任意のコンピュータで読み取り可能な記録媒体が用いられる。この可搬出記録媒体209に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ202にロードして使用することもできる。

【0184】ネットワーク接続装置207は、LAN、WAN等の任意のネットワーク(回線)を介して外部の装置と通信し、通信に伴うデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ202にロードして使用することもできる。

【0185】図19は、図18のコンピュータにプログラムとデータを供給することができる、コンピュータで読み取り可能な記録媒体及び伝送信号を説明する図である。上述のプログラムや各テーブルに格納されるデータを、以下のようにしてコンピュータ200に供給することにより、コンピュータ200に文書検索装置100に相当する機能を行なわせることも可能である。そのためには、上述のプログラムやデータを、コンピュータで読

み取り可能な記録媒体29に予め記憶させておく。そして、図19に示すように、媒体駆動装置206を用いて、記録媒体29からプログラム等をコンピュータ200に読み出させて該コンピュータ200のメモリ202や外部記憶装置205に一旦格納させ、そのコンピュータ200の有するCPU201にこの格納されたプログラムを読み出させて実行させるように構成すればよい。

【0186】また、記録媒体209からプログラムをコンピュータに読み出させる代わりに、プログラム(データ)提供者が有するDB210から、通信回線(ネットワーク)211を介して、プログラムをダウンロードすることとしてもよい。この場合、例えば、DB210を有しプログラムを送信するコンピュータでは、上記プログラムを表現するプログラム・データをプログラム・データ・シグナルに変換し、変換されたプログラム・データ・シグナルをモデムを用いて変調することにより伝送信号を得て、得られた伝送信号を通信回線211(伝送媒体)に出力する。プログラムを受信するコンピュータでは、受信した伝送信号をモデムを用いて復調することにより、プログラム・データ・シグナルを得て、得られたプログラム・データ・シグナルを変換することにより、プログラム・データを得る。

【0187】なお、送信側のコンピュータと受信側のコンピュータの間を接続する通信回線211(伝送媒体)がデジタル回線の場合、プログラム・データ・シグナルを通信することも可能である。また、データベース(DB)210を有し、プログラムを送信するコンピュータと、プログラムをダウンロードするコンピュータとの間に、電話局等のコンピュータが介在しても良い。

【0188】以上、本発明の実施形態について説明したが、本発明は上述した実施形態に限定されるものではなく、他の様々な変更が可能である。

(付記1) ネットワーク上の文書の人気の高さの度合いである人気度を算出する人気度算出方法であって、文書からリンク関係を抽出し、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する、ことを含むことを特徴とする人気度算出方法。

【0189】(付記2) 前記リンク関係及び前記文書の前記ネットワーク上の位置を示す文書位置情報に基づいて前記人気度を算出する、ことを更に含むことを特徴とする付記1に記載の人気度算出方法。

【0190】(付記3) 前記文書位置情報を示す文字列の特徴に基づいて、前記人気度を算出する、ことを更に含むことを特徴とする付記2に記載の人気度算出方法。

【0191】(付記4) 前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出する、ことを更に含むことを特徴とする付記1に記載の人気度算出方法。

【0192】(付記5) 第2の期間内に算出された前記人気度に基づいて、前記人気変化度を算出する、ことを更に含むことを特徴とする付記4に記載の人気度算出方法。

【0193】(付記6) 前記第2の期間内に算出された前記人気度の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出する、ことを更に含むことを特徴とする付記5に記載の人気度算出方法。

【0194】(付記7) 前記回帰式の回帰係数に基づいて前記人気変化度を決定する、ことを更に含むことを特徴とする付記6に記載の人気度算出方法。

(付記8) 前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定する、ことを更に含むことを特徴とする付記7に記載の人気度算出方法。

【0195】(付記9) 前記第2の期間内に算出された前記人気度に基づいて、前記抽出された文書中の各文書の順位を決定し、前記第2の期間内の前記順位の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出する、ことを更に含むことを特徴とする付記5に記載の人気度算出方法。

【0196】(付記10) ネットワーク上の文書間の関係を判定する文書関係判定方法であって、第1の文書からリンク関係を抽出し、前記リンク関係に基づいて、前記第1の文書からリンクされる第2の文書が、前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを含むことを特徴とする文書関係判定方法。

【0197】(付記11) 前記第1の文書から前記第2の文書にリンクする部分の近辺にある文字列を前記第1の文書から抽出し、前記文字列に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0198】(付記12) 前記文字列が特定の文字列である場合、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であると決定する、ことを更に含むことを特徴とする付記11に記載の文書関係判定方法。

【0199】(付記13) 前記第2の文書のファイル名の拡張子に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0200】(付記14) 前記拡張子が特定の拡張子でない場合、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書でないと決定する、ことを更に含むことを特徴とする付記13に記載の文書関係判定方法。

【0201】(付記15) 前記第2の文書が前記第1の文書内で所定回数以上使用されているか否かに基づい

て、前記第2の文書は前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0202】(付記16) 前記第2の文書が前記第1の文書内で所定回数以上使用されている場合、前記第2の文書は前記第1の文書の内容に関連する非テキスト文書でないと決定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0203】(付記17) 前記第2の文書が前記第1の文書内で所定回数以上使用されていない場合、前記第2の文書は前記第1の文書の内容に関連する非テキスト文書であると決定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0204】(付記18) 前記第1の文書内に前記第2の文書のファイル名と類似したファイル名を持つ第3の文書がある場合、前記第2の文書の前記ファイル名が前記第3の文書の前記ファイル名よりも辞書順に若くない場合、前記第2の文書を第1の文書の内容に関連する非テキスト文書としてデータベースに登録しない、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0205】(付記19) 前記第2の文書からリンクされる第3の文書があるか否か判定する、ことを更に含むことを特徴とする付記10に記載の文書関係判定方法。

【0206】(付記20) 前記第2の文書からリンクされる第3の文書がある場合、前記第1の文書の前記ネットワーク上の位置を示す文書位置情報と前記第2の文書の文書位置情報に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記19に記載の文書関係判定方法。

【0207】(付記21) 前記第1の文書の前記文書位置情報と前記第3の文書の文書位置情報に基づいて、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書であるか否か判定する、ことを更に含むことを特徴とする付記20に記載の文書関係判定方法。

【0208】(付記22) 前記第2の文書の前記文書位置情報と第3の文書の前記文書位置情報が、前記第1の文書の前記文書位置情報と同じサーバアドレス又はドメインを持たない場合、前記第2の文書が前記第1の文書の内容に関連する非テキスト文書でないと決定する、ことを更に含むことを特徴とする付記21に記載の文書関係判定方法。

【0209】(付記23) ネットワーク上の文書が提供するサービスの種別を判定するサービス種別判定方法であって、前記文書からユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定する、ことを含むことを特徴とするサービス種別判定方法。

【0210】(付記24) 前記文書に前記ユーザ入力に指定するタグが含まれていない場合、前記文書はサービスを提供しないと決定する、ことを更に含むことを特徴とする付記23に記載のサービス種別判定方法。

【0211】(付記25) 前記文書に含まれるボタンの表示に基づいて、前記文書が提供するサービスの種別を判定する、ことを更に含むことを特徴とする付記23に記載のサービス種別判定方法。

【0212】(付記26) 前記文書に含まれるユーザ入力エリアに基づいて、前記文書が提供するサービスの種別を判定する、ことを更に含むことを特徴とする付記25に記載のサービス種別判定方法。

【0213】(付記27) ネットワーク上の文書の人気の高さの度合いである人気度を算出する制御をコンピュータに実行させるプログラムであって、文書からリンク関係を抽出し、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する、ことを含む処理を前記コンピュータに実行させることを特徴とするプログラム。

【0214】(付記28) 前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出する、ことを更に含む処理を更にコンピュータに実行させることを特徴とする付記27に記載のプログラム。

【0215】(付記29) 第2の期間内に算出された前記人気度に基づいて、前記人気変化度を算出する、ことを更に含む処理を前記コンピュータに実行させることを特徴とする付記28に記載のプログラム。

【0216】(付記30) 前記第2の期間内に算出された前記人気度の時間に対する回帰式を算出し、前記人気変化度を前記回帰式に基づいて算出する、ことを更に含む処理を前記コンピュータに実行させることを特徴とする付記29に記載のプログラム。

【0217】(付記31) 前記回帰式の回帰係数に基づいて前記人気変化度を決定する、ことを更に含む処理を前記コンピュータに実行させることを特徴とする付記30に記載のプログラム。

【0218】(付記32) 前記回帰式の切片に基づいて、前記人気度の時間に対する推移の傾向を決定する、ことを更に含む処理を前記コンピュータに実行させることを特徴とする付記31に記載のプログラム。

【0219】(付記33) ネットワーク上の文書間の関係を判定する制御をコンピュータに実行させるプログラムであって、第1の文書からリンク関係を抽出し、前記リンク関係に基づいて、前記第1の文書からリンクされる第2の文書が、前記第1の文書の内容に関連する非テキストコンテンツであるか否かを判定する、ことを含む処理を前記コンピュータに実行させることを特徴とするプログラム。

【0220】(付記34) ネットワーク上の文書が提

供するサービスの種別を判定する制御をコンピュータに実行させるプログラムであって、前記文書からユーザ入力に指定するタグを抽出し、前記ユーザ入力に指定するタグに基づいて、前記文書が提供するサービスの種別を判定する、ことを含む処理を前記コンピュータに実行させることを特徴とするプログラム。

【0221】(付記35) ネットワーク上から文書を検索する文書検索方法であって、前記ネットワークから文書を収集し、前記文書からリンク関係を抽出し、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出し、検索条件に基づいて文書を検索し、前記検索された文書を前記人気度に基づいてランキングし、前記ランキング結果に基づいて、前記検索された文書に関する情報を出力する、ことを含むことを特徴とする文書検索方法。

【0222】(付記36) 第2の期間内に算出された前記人気度に基づいて、前記文書の前記人気度の変化の方向と度合いを示す人気変化度を算出し、前記人気変化度に関する情報を前記検索された文書に関連する情報に加える、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0223】(付記37) 前記リンク関係に基づいて、前記文書からリンクされる他の文書が、前記文書の内容に関連する関連非テキスト文書であるか否かを判定し、前記判定の結果に基づいて、前記関連非テキスト文書に関する情報を前記検索された文書に関連する情報に加える、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0224】(付記38) 前記関連非テキスト文書に関する情報に、前記関連非テキスト文書へのリンクを埋め込む、ことを更に含むことを特徴とする付記37に記載の文書検索方法。

【0225】(付記39) 前記文書からユーザ入力に指定するタグを抽出し、前記ユーザ入力に指定するタグに基づいて、前記文書が提供するサービスの種別を判定し、前記サービスの種別に関する情報を前記検索された文書に関連する情報に加える、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0226】(付記40) ユーザからある文書の前記ネットワーク上の位置を示す文書位置情報及び所定値の登録を受け付け、前記文書位置情報によって特定される前記文書の前記人気度が、前記所定値になった場合、前記人気度が前記所定値になった旨を前記ユーザに通知する、ことを更に含むことを特徴とする付記35に記載の文書検索方法。

【0227】(付記41) ネットワーク上から文書を検索する文書検索装置であって、前記ネットワークから文書を収集し、前記収集された文書からリンク関係を抽出する収集手段と、第1の期間内に更新又は収集された

文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する人気度算出手段と、検索条件に基づいて文書を検索し、前記検索された文書を前記人気度に基づいてランキングし、前記ランキング結果に基づいて、前記検索された文書に関する情報を出力する検索サービス手段と、を備えることを特徴とする文書検索装置。

【0228】(付記42) 地域に関する文書をネットワーク上から検索する地域情報文書検索装置であって、前記ネットワークから文書を収集し、前記収集された文書からリンク関係を抽出する収集手段と、第1の期間内に更新又は収集された文書を前記人気度を算出する対象として抽出し、前記抽出された各文書の人気度を算出する人気度算出手段と、第2の期間内に算出された前記人気度に基づいて、前記人気度の変化の方向と度合いを示す人気変化度を算出する人気度遷移算出手段と、前記収集された文書間のリンク関係に基づいて、各文書からリンクされる文書が、各文書の内容に関連する関連非テキスト文書であるか否かを判定する関連非テキストコンテンツ判定手段と、前記収集された文書からユーザ入力を指定するタグを抽出し、前記ユーザ入力を指定するタグに基づいて、前記文書が提供するサービスの種別を判定するサービス種別判定手段と、前記収集された文書を地域名毎に階層的に分類する分類手段と、ユーザから指定された地域名に基づいて文書を検索し、前記検索された文書を前記人気度に基づいてランキングし、前記ランキング結果に基づいて、前記検索された文書に関する情報とともに、前記検索された文書の前記人気変化度に関する情報、前記関連非テキスト文書に関する情報及び前記検索された文書が提供するサービス種別に関する情報を出力する検索サービス手段と、を備えることを特徴とする文書検索装置。

【0229】

【発明の効果】以上詳細に説明したように、本発明は、第1の期間内に収集又は更新された文書を対象として人気の高さの度合いを示す人気度を算出し、さらに、第2の期間内に算出された人気度に基づいて人気度の変化の度合いを示す人気変化度を算出する。これにより、文書の人気度が増加する一方で減少することがないという問題を解決しつつ、文書が時系列的にどのような状況にあるのかを示す情報を得る事を可能とする。

【0230】また、本発明によれば、文書間のリンク関係及びタグに基づいて、非テキストコンテンツ及びサービスを提供する文書等、多様な文書を整理する事が可能となる。

【図面の簡単な説明】

【図1】本発明の原理図である。

【図2】本発明に係わる文書検索装置の構成図である。

【図3】文書テーブルのデータ構造の一例を示す図である。

【図4】リンク関係テーブルのデータ構造の一例を示す図である。

【図5】人気度テーブルのデータ構造の一例を示す図である。

【図6】人気度変化テーブルのデータ構造の一例を示す図である。

【図7】非テキストコンテンツテーブルのデータ構造の一例を示す図である。

【図8】サービス種別テーブルのデータ構造の一例を示す図である。

【図9】人気度を算出する処理の手順を示すフローチャートである。

【図10】人気度の算出における本発明の特徴を説明する図である。

【図11】人気変化度を算出する処理の手順を示すフローチャートである。

【図12】関連する非テキストコンテンツを判定する処理の手順を示すフローチャートである。

【図13】提供するサービスを判定する処理の手順を示すフローチャートである。

【図14】検索結果の表示画面の一例を示す図である。

【図15】人気度推移画面の一例を示す図である。

【図16】本発明を適用した業界分析ツールの画面の一例を示す図である。

【図17】本発明を適用した地域情報検索システムの画面の一例を示す図である。

【図18】コンピュータの構成図である。

【図19】コンピュータにプログラムやデータを提供することができる記録媒体及び伝送信号を説明する図である。

【符号の説明】

10 文書整理装置

11 人気度算出手段

12 人気度遷移算出手段

13 関連非テキストコンテンツ判定手段

14 サービス種別判定手段

100 文書検索装置

101 収集部

102 人気度算出部

103 人気度遷移算出部

104 関連非テキストコンテンツ判定部

105 サービス種別判定部

106 ページ分類部

107 検索サービス部

108 ブラウザ

111 文書テーブル

112 リンク関係テーブル

113 人気度テーブル

114 人気度変化テーブル

115 非テキストコンテンツテーブル

116 サービス種別テーブル  
200 コンピュータ  
201 CPU  
202 メモリ  
203 入力装置  
204 出力装置  
205 外部記憶装置

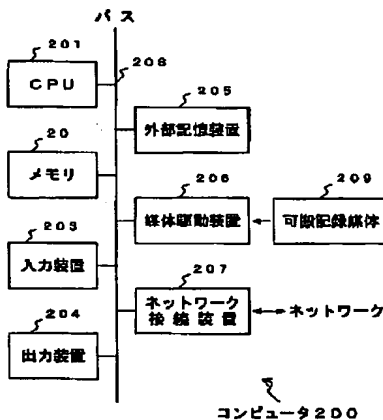
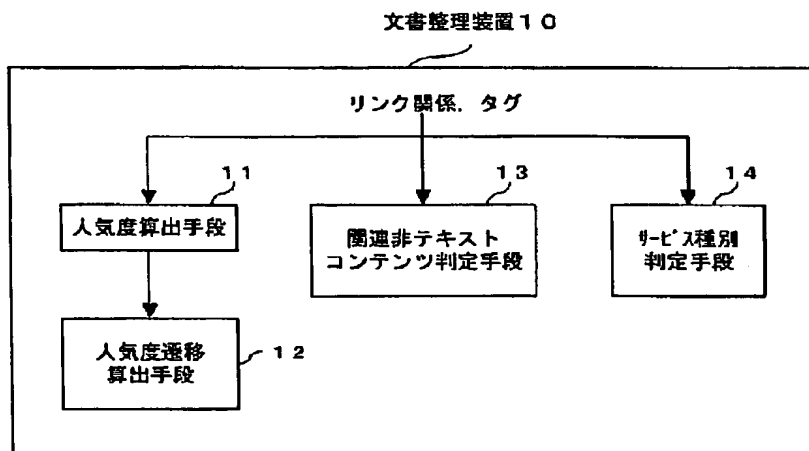
206 媒体駆動装置  
207 ネットワーク接続装置  
208 バス  
209 可搬記録媒体  
210 プログラム（データ）提供者  
211 回線

【図1】

【図18】

# 本 発 明 の 原 理 図

コンピュータの構成図



【図3】

【図7】

文書テーブルのデータ構造の一例を示す図 非テキストコンテンツテーブルのデータ構造の一例を示す図

文書テーブル111

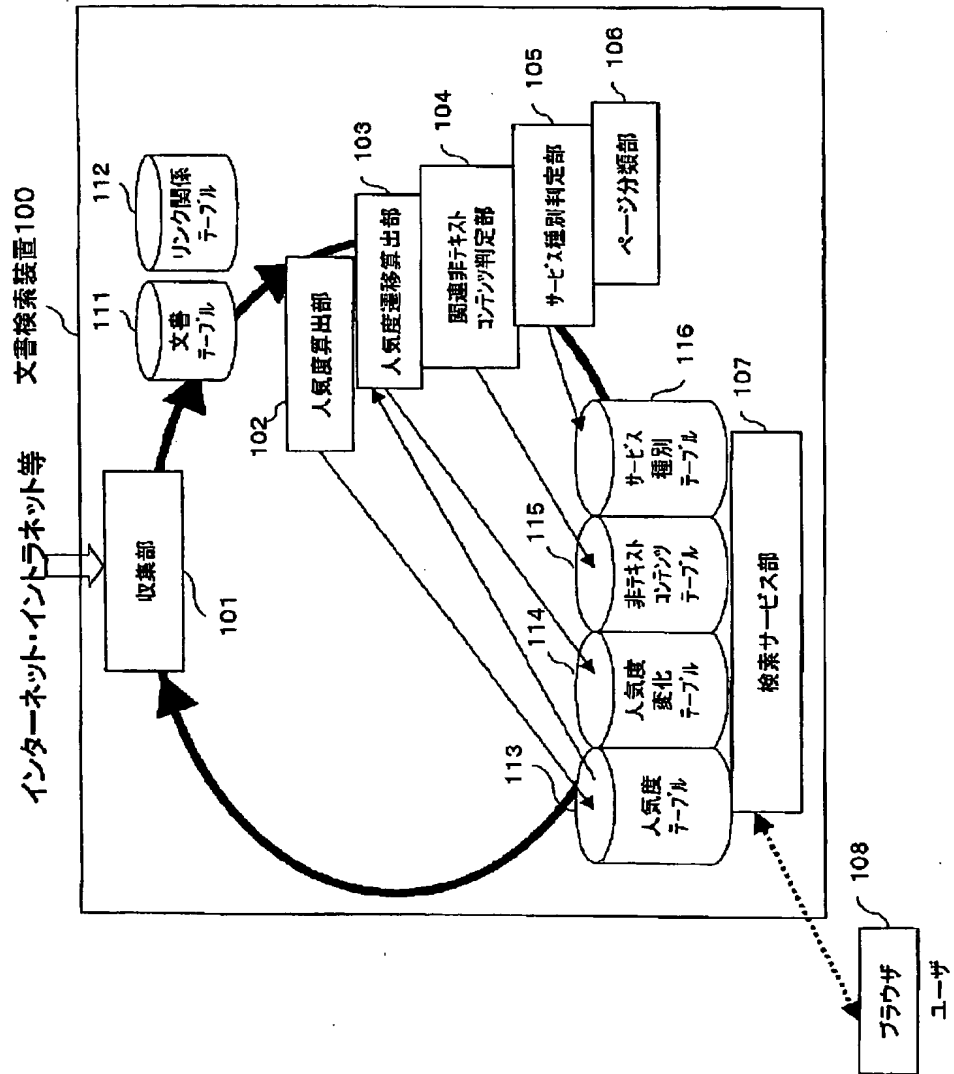
URL	ID
http://aaa.co.jp/	123
http://bbb.co.jp/dd/	124
...	...
...	...

非テキストコンテンツテーブル115

文書ID	関連非テキストコンテンツID	種別
123	3630	mv
123	3150	snd
...	...	...

【図2】

本発明に係わる文書検索装置の構成図



【 図 4 】

リンク関係テーブルのデータ構造の一例を示す図

リンク関係テーブル112

取集日	更新日	リンク元ID	リンク先ID列
010810	010725	123	124,128,3150,3630,...
010810	010620	124	256,975,1225,.....
....	....	....	...

【 図 5 】

人気度テーブルのデータ構造の一例を示す図

人気度テーブル113

算出日	文書ID	人気度	人気度順位
010820	123	5036	346
010820	124	83645	5890
....	...	...	....

【 図 6 】

【 図 8 】

サービス種別テーブルのデータ構造の一例を示す図

サービス種別テーブル118

文書ID	サービス種別
124	検索
123	ショップ

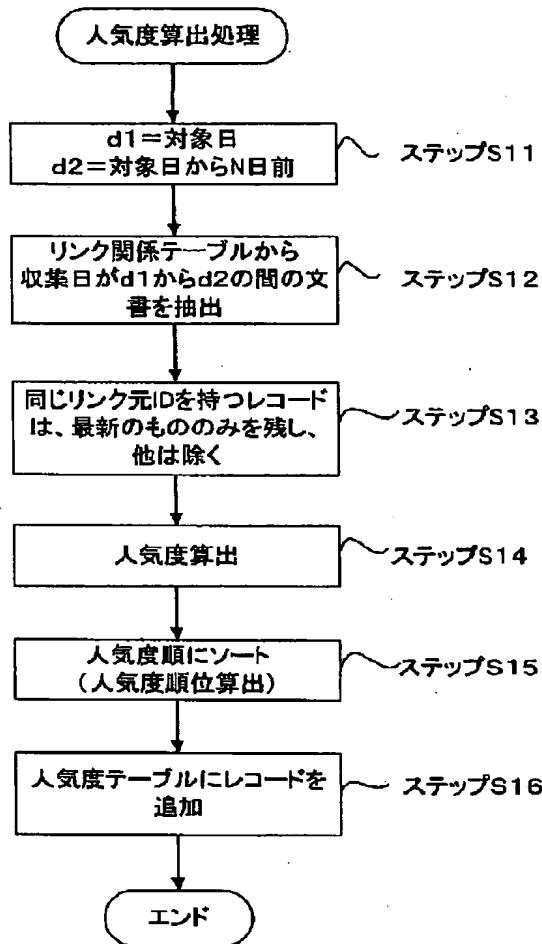
人気度変化テーブルのデータ構造の一例を示す図

人気度変化テーブル114

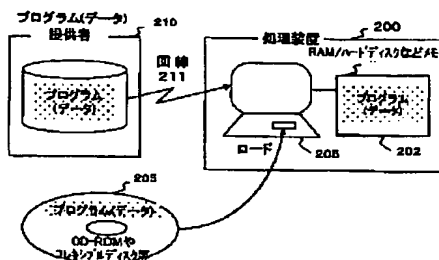
文書ID	人気度		人気度順位	
	傾き	切片	傾き	切片
123	-12	346	-6	233
124	-562	5890	-152	851
....	...	...		....

【図9】

人気度を算出する処理の手順を示すフローチャート

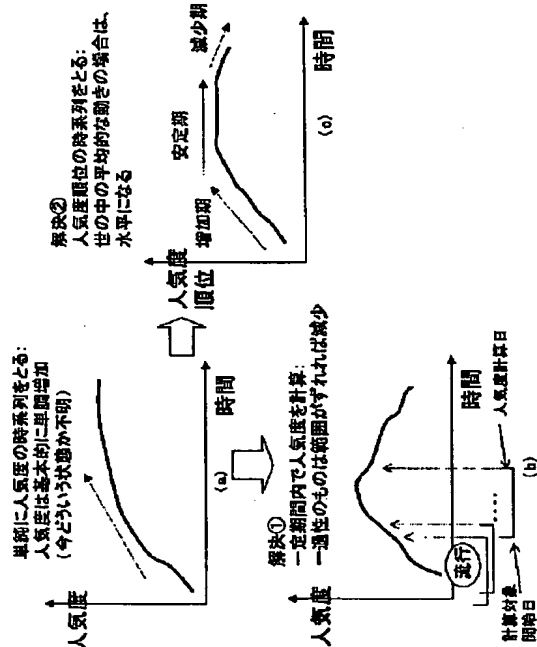


【図19】

コンピュータにプログラムやデータを提供することができる  
記憶媒体及び伝送信号を説明する図

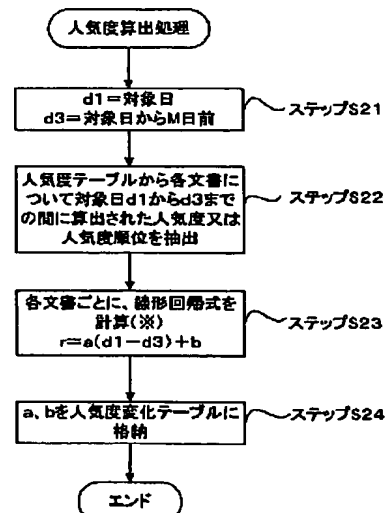
【図10】

人気度計算における本発明の特徴を説明する図



【図11】

人気変化度を算出する処理の手順を示すフローチャート



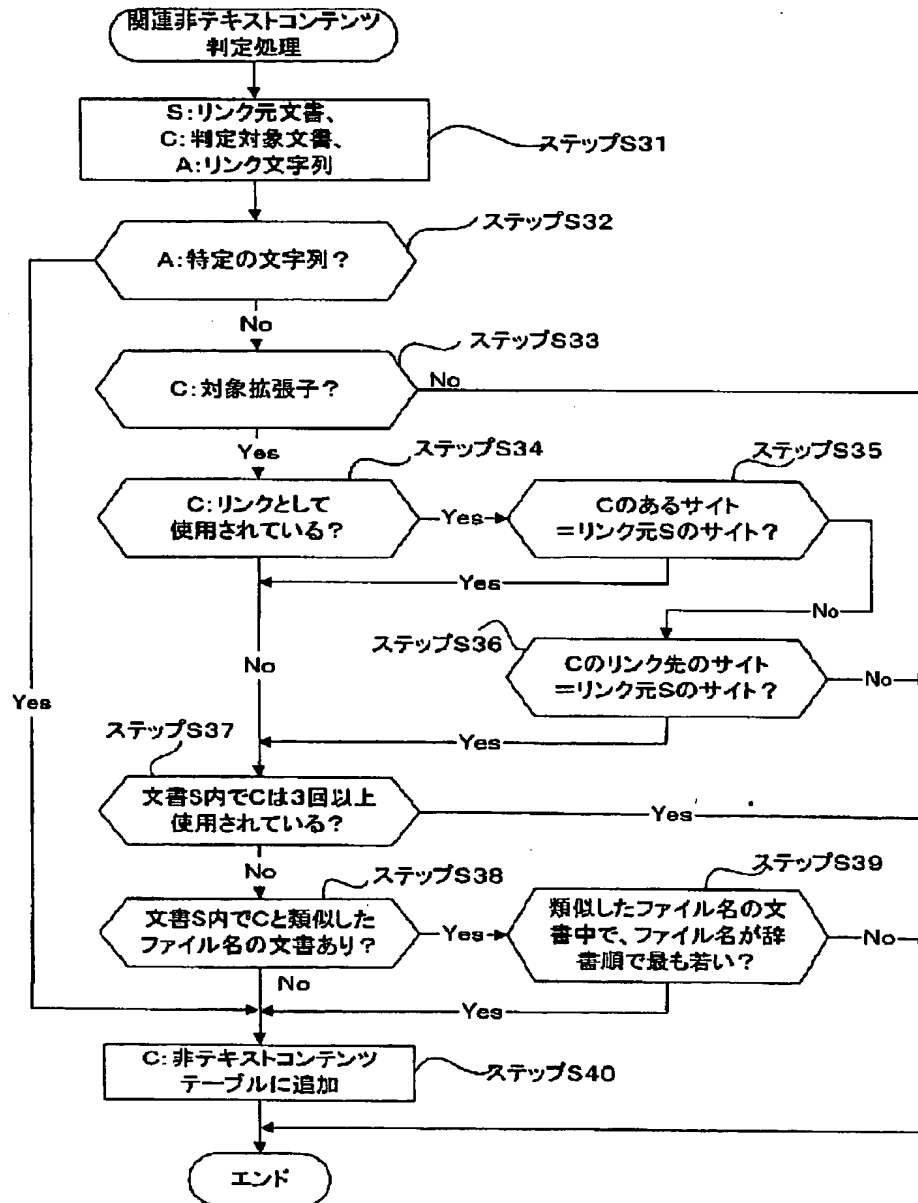
※d3, d3+1, ..., d1(M日)の人気度(順位)を $w_0, w_1, \dots, w_{M-1}$ とすると、  
 $a=(M \times lw - l \times W) / (M \times l^2 - l^2)$   
 $b=(l \times W - W \times l^2) / (l^2 - M \times l^2)$ である。  
 ただし、

$$lw = \sum_{i=0}^{M-1} i \cdot w_i, \quad W = \sum_{i=0}^{M-1} w_i, \quad l = \sum_{i=0}^{M-1} i, \quad l^2 = \sum_{i=0}^{M-1} i^2$$



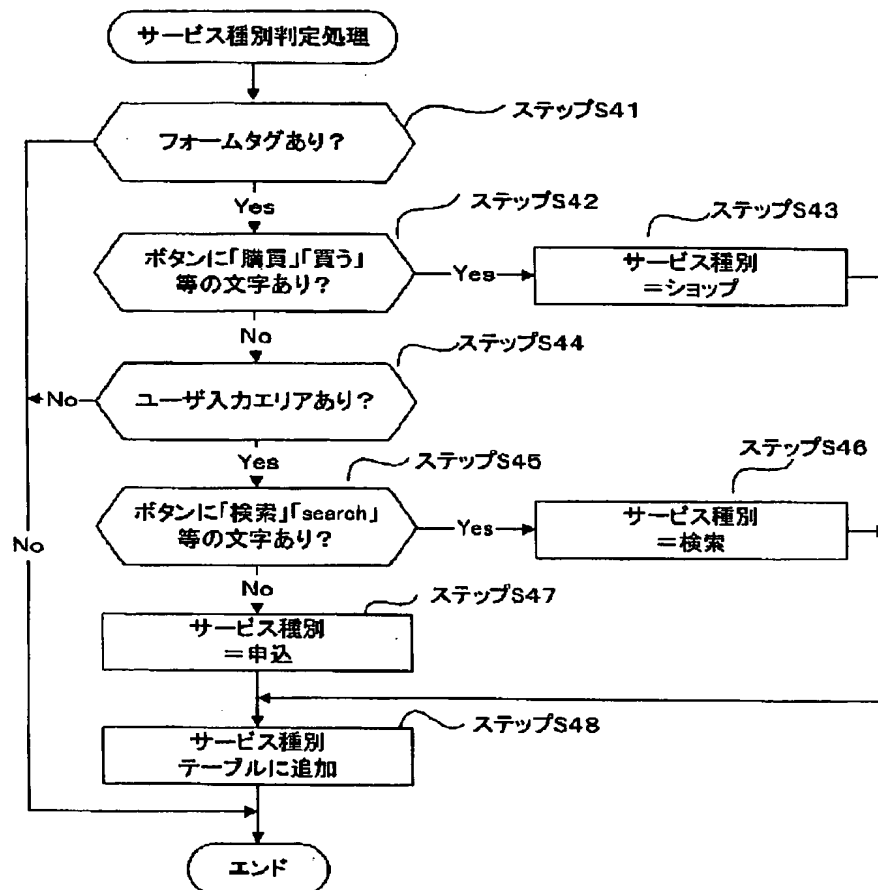
【図12】

関連する非テキストコンテンツを判定する処理の手順を示すフローチャート



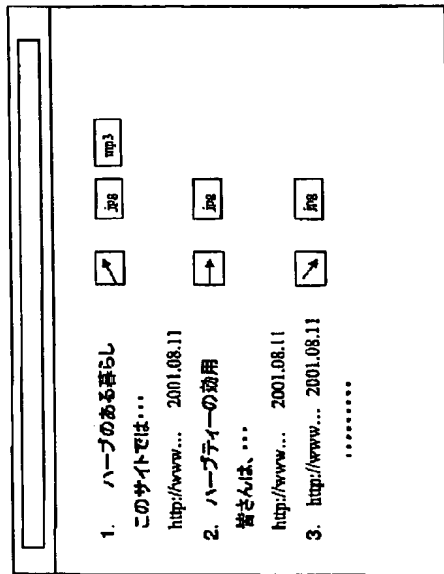
【図13】

提供するサービスを判定する処理の手順を示すフローチャート



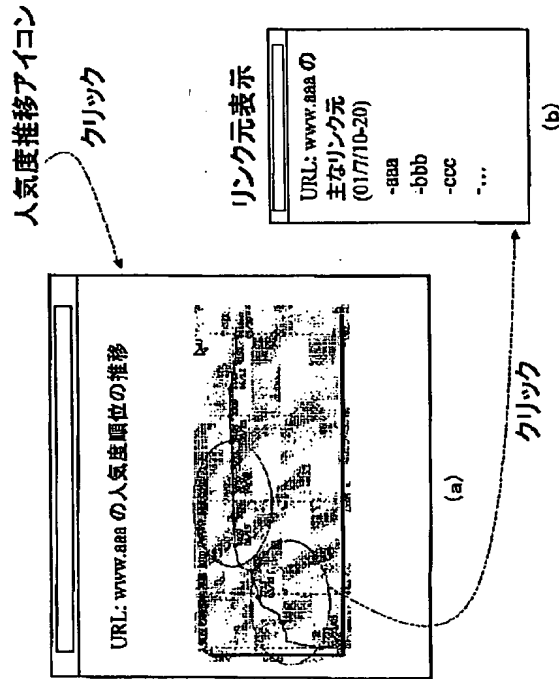
【図14】

検索結果の表示画面の一例を示す図



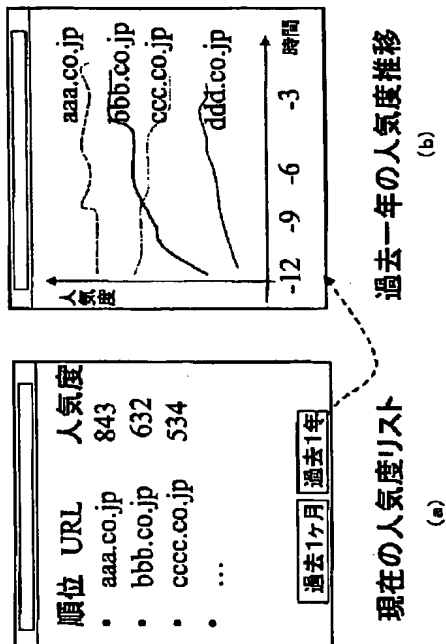
【図15】

人気度推移画面の一例を示す図



【図16】

本発明を適用した業界分析ツールの画面の一例を示す図



【図17】

本発明を適用した地域情報検索システムの画面の一例を示す図

